SKETCH – AN INVESTIGATION INTO FEATURE EXTRACTION IN COMPRESSED DOMAIN

KAZUKI MINEMURA

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

SKETCH – AN INVESTIGATION INTO FEATURE EXTRACTION IN COMPRESSED DOMAIN

KAZUKI MINEMURA

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Kazuki Minemura**

Registration/Matrix No.: **WHA120036**

Name of Degree: **Doctor of Philosophy**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Sketch – An Investigation into Feature Extraction in Compressed Domain

Field of Study: **Multimedia Signal Processing**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

The recent advancements of multimedia signal processing techniques enable us to conveniently edit/modify multimedia contents, which are usually available in the compressed form for transmission and storage purposes. Among the compression standards, JPEG for still image and H.264/AVC for video are the common formats that we handle in our daily life. As the application of compressed image/video widens, practical security tools for compressed image/video also become increasingly important. In that regard, encryption is the process designed to transform a content into an unintelligible form. Classical encryption approaches (e.g., DES, AES, RSA) handle the bitstream of a compressed image/video directly, but this approach invalidates many features of the compressed image/video, including hierarchical decoding and transcoding. In addition, the computational time of conventional encryption is also high. To overcome the aforementioned problems, the selective-encryption approach, which encrypts a subset of a compressed content, has been pursued by many researchers. Although many were proven to be robust against traditional cryptanalysis, the conventional selective-encryption methods are vulnerable to non-traditional forms of attack called sketch attack. In this thesis, five sketch attacks, i.e., four of them are based on DCT coefficients while another one exploits the statistical features of a coding block, are proposed. Their performances are evaluated and then compared. Furthermore, as a mean to evaluate the sketch image, an objective evaluation metric called outline clearness assessment (OCA) is put forward. Moreover, a selective encryption method for JPEG compressed image is put forward to address the problem of bitstream size overhead while being robust against both the traditional and non-traditional attacks. As an application, sketch attack is utilized as a feature extraction process in H.264/AVC compressed video for text detection purpose. Finally, contributions and limitations of this study are summarized, and future works are discussed.

ABSTRAK

Kemajuan kebelakangan ini dalam teknik pemprosesan isyarat multimedia memudahkan kita untuk mengubasuai/menyunting kandungan multimedia, yang biasanya didapati dalam bentuk termampat bagi tujuan penghantaran dan penyimpanan. Antara piawaian-piawaian mampatan, JPEG untuk imej and H.264/AVC untuk video adalah antara piawaian biasa yang kita menangani dalam kehidupan seharian. Sementara aplikasi imej/video termampat makin melebar, alat-alat keselamatan yang praktikal untuk imej/video termampat juga menjadi semakin penting. Sehubungan itu, penyulitan adalah proses yang direka untuk mengubah kandungan kepada bentuk yang tidak boleh difahami. Teknik-teknik penyulitan klasikal (contohnya, DES, AES, RSA) mengendalikan urutan bit daripada imej/video termampat secara langsung, tetapi teknik-teknik ini membatalkan banyak ciri-ciri imej/video termampat, termasuk penyahkodan hierarki dan transpengekodan. Di samping itu, kos pemprocessan penyulitan konvensional juga tinggi. Bagi mengatasi masalah-masalah yang dinyatakan di atas, teknik penyulitan selektif, yang menyulitkan subset kandungan termampat, telah diusahakan oleh para penyelidik. Walaupun banyak teknik tersebut telah dibukti kukuh terhadap kriptanalisis tradisional, teknik-teknik penyulitan selektif konvensional adalah lemah terhadap serangan bukan tradisional yang dipanggil Sketch Attack. Dalam tesis ini, lima Sketch Attack, iaitu, empat daripada mereka adalah berdasarkan pekali DCT manakala satu lagi mengeksploitasikan ciri-ciri statistik blok pengekod, telah dicadangkan. Pencapaian mereka telah dinilai dan dibanding. Selain itu, sebagai cara untuk menilai Sketch Attack, objektif penilaian metrik yang dikenali sebagai Outline Clearness Assessment (OCA) telah dicadangkan. Di samping itu, kaedah penyulitan selektif untuk imej termampat dalam piawaian JPEG telah dicadangkan untuk menangani masalah perkembangan saiz urutan bit pada masa yang sama kukuh terhadap kedua-dua serangan tradisional dan bukan tradisional. Sebagai aplikasi,

Sketch Attack digunakan sebagai proses pengekstrakan ciri-ciri dari video termampat dalam piawaian H.264/AVC untuk tujuan pengesanan teks. Akhir sekali, sumbangan dan kekurangan kajian ini diringkaskan, dan kerja-kerja masa depan dibincangkan.

ACKNOWLEDGEMENTS

First of all, I would like to express that I have great appreciation to my supervisor Dr. Kok-Sheik Wong for his invaluable support and guidance from the first moment I started this research work to the moment I finished the work. I have learned invaluable lessons from his invaluable comments and discussion in various ways. Next, I would like to greatly thank MSPIH group members for great research and non-research activities, which greatly have helped my Ph.D. candidate life. I have really enjoyed the constructive discussions with Dr. Wong and MSPIH members throughout the preparation of this study, which enriched me in the spirit of patience, discipline and hard-working. Not to forget my financial support, High Impact Research Grant. It is the University of Malaya projects, which will result in the publication of manuscripts in Tier 1 ISI/Web of Science journals. I'm glad and feel honored to be a part of this great project. Last but not least, I would like to express the great gratitude for my parents warm understanding and patient support to my research career.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

D_1	:	Uncompressed domain.
D_2	:	Compressed domain.
D_3	:	Bitstream domain.
E	:	The expectation function.
$F-measure$	:	$2 \times Precision \times Recall/(Precision + Recall)$.
F_1	:	Format-compliant.
F_2	:	Non format-compliant.
M	:	The vertical number of blocks.
N	:	The horizontal number of blocks.
$Precision$	:	Number of correct lines over number of collected lines.
$QT[u][v]$	:	The (u,v)-th quantization value.
$Recall$	:	Number of correct lines over ground truth.
T_1	:	Full encryption.
T_2	:	Selective encryption.
X	:	The height of the image $g(x,y)$.
Y	:	The width of the image $g(x,y)$.
Δ	:	The number of 0's in Canny edge image Γ.
Δ_0	:	The number of 0's in Sobel edge image Γ_0.
Γ	:	The Canny edge image.
Γ_0	:	The Sobel edge image.
Γ_D	:	The down-scaled Canny binary edge map.
Γ_N	:	The number of edge candidates for each block.
Λ	:	Information as entropy.
Λ_f	:	The global feature of an image g.
Λ_g	:	GIIE.
Λ_t	:	A threshold value.
Ω_R	:	The refined text regions.
Θ	:	The number of 1's in Canny edge image Γ.
Θ_0	:	The number of 1's in Sobel edge image Γ_0.
Ξ_{bl}	:	SSIM between original and uBL.
Ξ_{br}	:	SSIM between original and uBR.
Ξ_m	:	The local feature of an image g.
Ξ_{tl}	:	SSIM between original and uTL.
Ξ_{tr}	:	SSIM between original and uTR.
α	:	Row index in a table.
β	:	Column index in a table.
χ	:	The number of AC sign information.
δ	:	The size of a block-transform or MB.

η_1	:	The number of $\Gamma_0(x,y) = 0$ & $\phi(x,y) = 0$.
$\eta_1^{'}$	:	The number of $\Gamma(i,j) = 0$ & $\phi(i,j) = 0$.
η_2	:	The number of $\Gamma_0(x,y) = 1$ & $\phi(x,y) = 1$.
$\eta_2^{'}$	:	The number of $\Gamma(i,j) = 1$ & $\phi(i,j) = 1$.
η_3	:	The number of $\Gamma_0(x,y) = 0$ & $\phi(x,y) = 1$.
$\eta_3^{'}$	:	The number of $\Gamma(i,j) = 0$ & $\phi(i,j) = 1$.
η_4	:	The number of $\Gamma_0(x,y) = 1$ & $\phi(x,y) = 0$.
$\eta_4^{'}$	:	The number of $\Gamma(i,j) = 1$ & $\phi(i,j) = 0$.
γ	:	Transform window size $\in 4, 8$.
κ	:	The number of categories.
$\lambda_C(i,j)$	:	The sum of absolute complex AC coefficients in (i,j)-th block.
$\lambda_D(i,j)$	:	The sum of absolute diagonal AC coefficients in (i,j)-th block.
$\lambda_V(i,j)$	:	The sum of absolute vertical AC coefficients in (i,j)-th block.
$\mathbb{A}_{u,v}(i,j)$	:	The absolute (u,v)-th transformed coefficient of the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$.
$\mathbb{D}_1$	:	Inverse RDC.
$\mathbb{D}_2$	:	DC prediction error decryption.
$\mathbb{D}_3$	:	AC sign reconstruction.
$\mathbb{D}_4$	:	AC block reconstruction.
$\mathbb{D}_5$	:	AC ZRV pair decryption.
$\mathbb{F}(i,j)$	:	The (i,j)-th block in $\mathbb{T}$.
$\mathbb{F}^{'}(i,j)$	:	The arranged (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$.
$\mathbb{F}_{AC}(i,j)$	:	The (i,j)-th AC block, i.e., $\mathbb{F}(i,j)$ without the DC component.
$\mathbb{F}_{u,v}(i,j)$	:	The (u,v)-th transformed and quantized coefficient value of the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$.
$\mathbb{F}_{u,v}^{'}(i,j)$	:	The shuffled (u,v)-th transformed coefficient of the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$.
$\mathbb{J}_1$	:	RDC.
$\mathbb{J}_2$	:	DC error category mapping.
$\mathbb{J}_3$	:	AC sign randomization.
$\mathbb{J}_4$	:	AC block shuffling.
$\mathbb{J}_5$	:	AC ZRV pair shuffling.
$\mathbb{M}$	:	The vertical size of an outline image.
$\mathbb{N}$	:	The horizontal size of an outline image.
$\mathbb{O}$	:	The set of pixel values in g.
$\mathbb{P}$	:	The set of pixel values in g_α.
$\mathbb{T}$	:	The transformed coefficient values.
$\mathbb{T}_{u,v}(i,j)$	:	The (u,v)-th transformed coefficient value of the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$.

$\mathbb{Z}$	:	A random variable.
$\mu_{\mathbb{O}}$	:	The average pixel value of the input image g.
$\mu_{\mathbb{P}}$	:	The average of the shifted image g_α.
$\omega_{\mathbb{PO}}$	:	The spatial correlation value.
ϕ	:	Sketch image.
ϕ_B	:	MBS image.
ϕ_D	:	DCEC image.
ϕ_E	:	EAC image.
ϕ_N	:	INCC image.
ϕ_P	:	PLZ image.
ϕ_S	:	SAC image.
ϕ_T	:	The fused image.
ϕ_Z	:	NZCA image.
$\phi_Z^{'}(i,j)$	:	The pseudo-luminance value in (i,j)-th $\gamma \times \gamma$ block.
ϕ_{com}	:	The complex entry.
ϕ_{dia}	:	The diagonal entry.
ϕ_{hor}	:	The horizontal entry.
ϕ_{ver}	:	The vertical entry.
$\rho(l)$	:	The probability of gray level l.
$\rho(z)$	:	The probability of a random value z.
σ_A	:	The standard deviation value in AGWN.
$\sigma_{\mathbb{O}}$	:	The variance of the pixel values in g.
$\sigma_{\mathbb{P}}$	:	The variance of the shifted image g_α.
$\tau(\mathbb{J}_1)$	:	The number of combinations for region labels.
$\tau(\mathbb{J}_2)$	:	The number of combinations for DC errors categories.
$\tau(\mathbb{J}_3)$	:	The number of combinations for AC sign information.
$\tau(\mathbb{J}_4)$	:	The number of combinations for ZRV pairs.
$\tau(\mathbb{J}_5)$	:	The number of combinations for block shuffling.
$\tau(\mathbb{J}_l)$	:	The number of combinations for $\mathbb{J}_l$ encryption method.
Level_{AC}	:	The quantized AC coefficient values.
Level_{DC}	:	Residual DC values.
θ	:	The number of regions deduced from the AC coefficients.
ξ	:	Outline clearness score.
ζ	:	Absolute ESS equation.
ζ_0	:	ESS equation.
$constant$	:	A constant value.
$d(i,j)$	:	The category of $r(i,j)$.

$e(i,j)$	:	The strength of the edges/textures in the (i,j)-th 8×8 block.
g	:	An image.
$g(x,y)$	:	The $(x,y)-th$ pixel value in the image/frame g.
g_α	:	The values of *hor*, *ver*, *dig*.
$g_{\text{dia}}(x,y)$	:	The values of $g(x+1,y+1)$.
$g_{\text{hor}}(x,y)$	:	The values of $g(x,y+1)$.
$g_{\text{ver}}(x,y)$	:	The values of $g(x+1,y)$.
i	:	The vertical block index $\in [1,M]$.
i'	:	An integer value.
j	:	The horizontal block index $\in [1,N]$.
j'	:	An integer value.
k_1	:	A secret key.
k_2	:	A secret key.
k_3	:	A secret key.
k_4	:	A secret key.
l	:	The gray level.
$n(i,j)$	:	The number of nonzero AC coefficients in the (i,j)-th $\gamma \times \gamma$ block.
n_1	:	A NZCA threshold value.
n_2	:	A NZCA threshold value.
$p(i,j)$	:	The position of the last nonzero AC coefficients in the (i,j)-th 4×4 block.
$r(i,j)$	:	The residue DC value (i.e., prediction error) at the (i,j)-th 4×4 block.
$s(i,j)$	:	The sum of absolute of AC coefficients in the (i,j)-th 4×4 block.
u	:	The vertical coefficient index $\in [1,\gamma]$.
u'	:	An integer value.
v	:	The horizontal coefficient index $\in [1,\gamma]$.
v'	:	An integer value.
x	:	The vertical index $\in [1,X]$.
y	:	The horizontal index $\in [1,Y]$.
z	:	A random variable value.
AC	:	Alternating Current.
ACBS	:	AC Block Shuffling.
ACPS	:	AC ZRV Pair Scrambling.
ACSR	:	AC Sign Randomization.
ACSS	:	AC Subband Shuffling.
AES	:	Advanced Encryption Standard.
AGWN	:	Additive Gaussian White Noise.
ARISM	:	Autoregressive-Based Image Sharpness Metric.
AVS	:	Audio Video Standard.

B-	:	Bi-Predictive-.
BL	:	Bottom-Left.
BP	:	Baseline Profile.
BPH	:	Bitstream Packet Header.
BR	:	Bottom-Right.
BS	:	Block Shuffling.
BSDS500	:	Berkeley Segmentation Dataset and Benchmark.
CABAC	:	Context Adaptive Binary Arithmetic Coding.
CAN	:	Canny Edge Detector.
CAVLC	:	Context Adaptive Valuable Length Coding.
Coef_{no}	:	The number of nonzero AC coefficients of each block.
Coef_{po}	:	The position of the last nonzero AC coefficient in the order of Zigzag in each block.
CVM	:	Coefficient Value Mapping.
dB	:	Decibel.
DC	:	Direct Current.
DCEC	:	DC Error Category Attack.
DCEM	:	DC Error Mapping.
DCS	:	DC Shuffling.
DCSR	:	DC Sign Randomization.
DCT	:	Discrete Cosine Transform.
DES	:	Data Encryption Standard.
DF	:	Deblocking Filter.
DFV	:	Deblocking Filter Value Modification.
DMOS	:	Difference Mean Opinion Score.
DMV	:	Differential Motion Vector.
DPCM	:	Differential Pulse Code Modulation.
EAC	:	Energy of AC Coefficient Attack.
EC	:	Entropy Code.
ESS	:	Edge Similarity Score.
F.B.	:	Fishing Boat.
FMO	:	Flexible MB Ordering.
GIIE	:	Grayscale Image Information Entropy.
H.264/5	:	H.264 and H.265.
H.264/AVC	:	H.264 Advance Video Coding.
HD	:	High Definition.
HEVC	:	High Efficiency Video Coding.
I-	:	INTRA-.
ICDAR2013	:	International Conference Document Analysis and Recognition 2013.
INCC	:	Improved Nonzero Coefficient Count Attack.
IntDCT	:	Integer Discrete Cosine Transform.

INTER-M	:	INTER Prediction Mode.
INTER-MM	:	INTER-Prediction Mode Modification.
INTRA-M	:	INTRA Prediction Mode.
INTRA-MM	:	INTRA-Prediction Mode Modification.
IQA	:	Image Quality Assessment.
ISO/ITU	:	International Standard Organization and International Telecommunication Union.
JPEG	:	Joint Picture Expert Group.
JPEG-XR	:	JPEG Extend Range.
JPEG2000	:	Joint Picture Expert Group 2000.
L	:	IntDCT Coefficient Level Modification.
LIVE	:	Laboratory for Image and Video Engineering.
LOG	:	Laplacian of Gaussian.
m	:	Minute.
MB	:	Macro Block.
MB bits	:	The number of bits allocated to each MB.
MBAmap	:	MB Allocation MAP.
MBS	:	MB Bitstream Size Attack.
MBT	:	MB Type.
MP	:	Megapixel.
MPEG	:	Moving Picture Experts Group.
MSGF-PR	:	Multi-domain Structural and Global Frequency Features + Piecewise Regression.
MV	:	Motion Vector.
MVD	:	MV Difference Modification.
N	:	North.
NALU	:	Network Abstract Layer Unit.
NCC	:	Nonzero Coefficient Count Attack.
NE	:	North-East.
NFERM	:	Non-Reference Free Energy-Based Robust Metric.
NW	:	North-West.
NZCA	:	Nonzero Coefficient Count Attack.
OCA	:	Outline Clearness Assessment.
P-	:	Predictive-.
PCM	:	Pulse-Code Modulation.
PLZ	:	Position of the Last Nonzero Coefficient Attack.
PM	:	Prediction Mode.
PSNR	:	Pixel Signal-to-Noise Ratio.
QC	:	Quality Control Parameter.
QF	:	Quality Factor.
QP	:	Quality Parameter.
QPM	:	Quantization Parameter Manipulation.

QT	:	Quantization Table.
RDC	:	Rearranging DC Coefficients.
RGB	:	Red-Green-Blue.
round $(\cdot)$	:	Rounding the real value to an 8-bit unsigned integer in the range of $[0, 255]$.
RSA	:	Rivest-Shamir-Adleman.
S	:	IntDCT Coefficient Sign Randomization.
s	:	Second.
S.L.	:	Sailboat on Lake.
S3IM	:	Subsample Based SSIM.
SAC	:	Sum of Absolute AC Coefficients Attack.
SFE	:	Structure Forest Based Edge Detector.
SNS	:	Social Network Service.
SOB	:	Sobel Edge Detector.
SS	:	Surveillance System.
SSEQ	:	Spatial-Spectral Entropy-Based Quality.
SSIM	:	Structural Similarity Index Measurement.
SSO	:	Secret Scan Order.
ST	:	Secret Transformation.
subMB	:	Sub MB.
TL	:	Top-Left.
TR	:	Top-Right.
uBL	:	The upsampled images of BL pixels.
uBR	:	The upsampled images of BR pixels.
USC-SIPI	:	University of Southern California - Signal and Image Processing Institute.
uTL	:	The upsampled images of TL pixels.
uTR	:	The upsampled images of TR pixels.
VA	:	Video Advertisement.
VC	:	Video Conferencing.
VLC	:	Valuable Length Coding.
VOD	:	Video on Demand.
VSS	:	Video Sharing Service.
W	:	West.
YCbCr	:	Luminance, Chroma-Blue Chroma-Red.
ZRV	:	Zero-Run Value.

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview

An introduction of this study is presented in this chapter. It includes the overview of feature extraction, problem statements, aims and objectives, research methodology, scopes and limitations, and contributions, under the general topic of feature extraction in the compressed domain.

1.2 Overview on Feature Extraction in the Compressed Domain

The recent growth of computer networks and the advent of affordable devices enable us to utilize multimedia sharing services, Social Network Service (SNS) (e.g., Facebook, Twitter and Instagram), Video on Demand (VOD) (e.g., Netflix), Video Sharing Service (VSS) (e.g., Youtube), Video Conferencing (VC) (e.g., Google Hangouts), Video Advertisement (VA), Surveillance System (SS), etc. As a result, various digital image/video contents, including high-resolution images and High Definition (HD) videos, started to play crucial roles in our daily life. Generally, these images/videos are stored in widely deployed compression standards, namely, JPEG or H.264/AVC. Therefore, practical security tools for compressed images/videos become important.

In the literature, there is a practical class of encryption methods, namely, format-compliant selective encryption, which is satisfying both conventional cryptanalysis attacks and perceptual attacks. However, Li, et at. (W. Li & Yuan, 2007) put forward a new class of perceptual attack, which utilizes a primitive signal processing operation, where a set of coarse outlines can be obtained directly from the bitstream of an image encrypted in JPEG compressed domain. Since block-transform is one of the core parts in both JPEG and H.264/AVC, a simple single feature extraction is viable to sketch an outline of the encrypted images/videos, which are perceptually degraded by conventional

format-compliant selective encryption methods detailed in Chapter 2. However, this research direction has not been explored much. Literature survey performed during this study reveals that Li et al. (W. Li & Yuan, 2007) is the only work that sketches outline directly from the encrypted image. Hence, there is much room to design this class feature extractions for JPEG and H.264/AVC. Hereinafter, this class of feature extraction techniques and the extracted features are referred to as *sketch attacks* and *sketch images*, respectively.

In addition, Li et al. (W. Li & Yuan, 2007) evaluated the sketch images by referring to an edge image obtained from the original image/video. However, the original image cannot be obtained when only encrypted images/videos are considered, which is usually the case. An intuitive alternative assessment is to perform subjective evaluation by humans. Such subjective evaluations are not only slow, cumbersome, time-consuming but also expensive. Hence, assessment is also one of the challenging problems in feature extraction in the compressed domain.

Furthermore, other applications in the compressed domain, including text detection (X. Qian et al., 2012) and human action recognition (Tom et al., 2015), have been put forward recently. This fact suggests that researchers have given some attentions to the compressed domain, where sketch images obtained by sketch attacks may be useful for applications based on feature extraction in the compressed domain. Therefore, the demonstration of applications by using sketch images is significant for the investigation into feature extraction in the compressed domain. In that regards, encryption for block-transform compressed images and text detection in compressed videos are demonstrated in this study.

1.3 Problem Statements

As pointed out in the overview above, there are some problems to be solved. Existing format-compliant selective encryption may not be robust to sketch attack (W. Li & Yuan, 2007) and there are many opportunities to design sketch attacks, because there are many unprocessed components, such as coefficient category, coefficient energy, coefficient distribution, allocated bits to Macro Block (MB), etc., in block-transform based compression, e.g., JPEG and H.264/AVC.

In addition, Li et al. (W. Li & Yuan, 2007) considered only binary information for sketch image assessment, and manual thresholding is required. The utilization of the original image is also impractical because the plaintext image is usually unavailable. Hence, no-reference assessment for sketch images is desired.

Furthermore, features in the compressed domain have been exploited not only for the purpose of attacking encryption methods (W. Li & Yuan, 2007), but also for other applications (X. Qian et al., 2012; Tom et al., 2015). However, even the latest study of the feature extractions and its related applications in the compressed domain are still immature, because Qian et al. (X. Qian et al., 2012) only verified the performance of text detection for graphic text in INTRA- (I-)frame of H.264/AVC compressed video, while a compressed video is, by far, dominated by Predictive- (P-) or Bi-Predictive- (B-)frames. On the other hand, although Tom et al. (Tom et al., 2015) utilized Quality Parameter (QP) and Motion Vector (MV) as features for human action recognition, there are still many unexplored features within the H.264/AVC standard.

Based on the descriptions above, the current problems in the compressed domain are summarized as follows:

1. Existing format-compliant selective encryption may not be robust against sketch attack.

2. There is still much room for improvement in designing sketch attack for JPEG images and H.264/AVC videos.

3. The literature is lacking of no-reference assessment of sketch/outline images.

4. The limited scope of exploration for feature extraction in the compressed domain.

1.4 Aims and Objectives

This study aims at investigating into feature extractions in the compressed domain for purposes of managing digital images/videos. To achieve this goal, this study aims to achieve the following objectives:

1. To conduct a comprehensive survey of feature extraction in the compressed domain;

2. To develop novel sketch attacks for encrypted image and video;

3. To develop a no-reference assessment metric for sketch image;

4. To develop a robust format-compliant selective encryption framework;

5. To evaluate the newly proposed sketch attacks and encryption technique, and;

6. To develop a text detection in the compressed domain as an application of sketch attack.

1.5 Research Methodology

To achieve these objectives, the following five-step methodology is considered in this thesis, including: literature survey; sketch attacks for format-compliant selectively encrypted video; outline clearness assessment; format-compliant selective encryption framework for block-transform compressed image, and text detection in H.264/AVC compressed video. First, literature survey is conducted to study the components of JPEG image and H.264/AVC video, and encryption. Next, the problems related to feature extraction in

the compressed domain and the research objectives are identified. Then, sketch attacks for video encrypted by using format-compliant selective encryption techniques are proposed and evaluated. Furthermore, a no-reference outline clearness assessment metric is proposed for giving objective scores to sketch images. Moreover, a format-compliant selective encryption framework is put forward to address the current encryption problems. Finally, text detection using sketch attack in the compressed domain is demonstrated as one possible applications.

1.6 Scopes and Limitations

The scope and limitations of this study need to be addressed concisely in order to layout an achievable research plan and proceed according to schedule for achieving the research objectives. The scopes and the limitations are as follows:

1. The target contents in this study are limited to block-transform compressed images/videos, considering the state-of-the-art compression standards, namely, JPEG and H.264/AVC.

2. This study considers both non-encrypted (i.e., plaintext) contents and encrypted contents attained by format-compliant selective encryption methods, which can flexibly generate images/videos of desired levels of distortion.

3. International Standard Organization and International Telecommunication Union (ISO/ITU) reference software under C programming language are utilized to generate compressed as well as compressed-encrypted images and videos.

1.7 Contributions

This study has complied with comprehensive works, which are translated to the contributions below:

1. Novel sketch attacks are designed to generate sketch images directly from both INTRA- and INTER-frames of compressed-only as well as compressed-encrypted videos.

2. A clearness assessment metric for outline images (sketch images and edge images) is formulated.

3. A novel format-compliant selective encryption framework for block-transform compressed images is put forward to achieve encryption, which is robust against both traditional cryptanalyses as well as sketch attacks.

4. A text detection method is proposed to demonstrate the viability of sketch image in other domain, in addition to its ability in attacking format-compliant encryption methods.

1.8 Structure of Thesis

This thesis is organized as follows. In Chapter 2, commonly utilized block-transform compression image/video standards, conventional format-compliant selective encryption methods, attacking methods (viz., cryptographic attacks, perceptual attacks and a sketch attack) and a conventional sketch image assessment metric in the literature are surveyed. Then, the problem of feature extraction directly from encrypted images/videos is detailed. In Chapter 3, a novel feature extraction method and a sketch attack framework are proposed. To evaluate sketched images, a non-reference outline assessment metric is proposed to evaluate the quality of the sketch and edge images in Chapter 4. To overcome sketch attack, a novel scrambling framework for block-transform compressed images is proposed in Chapter 5. The viability of the extracted features in other applications is further investigated, where, text detection in the compressed domain is demonstrated in Chapter 6. Finally, in Chapter 7, the conclusions and future directions of this study are

discussed.

CHAPTER 2: LITERATURE SURVEY

2.1 Overview

In this chapter, literature related to feature extraction in the compressed domain is overviewed as following topics: compression standards for still image and video; format-compliant encryption; sketch attack; quality assessment, and; applications in the compressed domain. Specifically, the components of two compression standards, i.e., JPEG and H.264/AVC, are scrutinized. Then, format-compliant selective encryption methods for images/videos and attacking methods are described. Next, a conventional sketch attack is reviewed, followed by an overview of sketch image quality assessment for edge image. In addition, an overview of text detection in the spatial domain and the compressed domain are also presented. Finally, the problems in the surveyed literature are identified and discussed.

2.2 Overview of block-transform Compression Standards

In this study, two widely deployed compression standards are considered throughout this study, namely, JPEG and H.264/AVC.

2.2.1 JPEG

In spite of the fact that JPEG (ISO/IEC, 1994) compression was standardized a few decades ago, the format is adopted in most of the decoders, which are included in digital devices (such as smartphone, laptop), and the format is utilized as a still compressed image file format in digital cameras. Hence, the number of images sorted in JPEG format is huge.

Figure 2.1 shows the main components in the JPEG compression standard. First, the level shift operation is applied to the intensity values of the original image to shift them by -128, and then the output is converted from Red-Green-Blue (RGB) to Luminance, Chroma-Blue Chroma-Red (YCbCr) (i.e., color conversion operation). Here, the

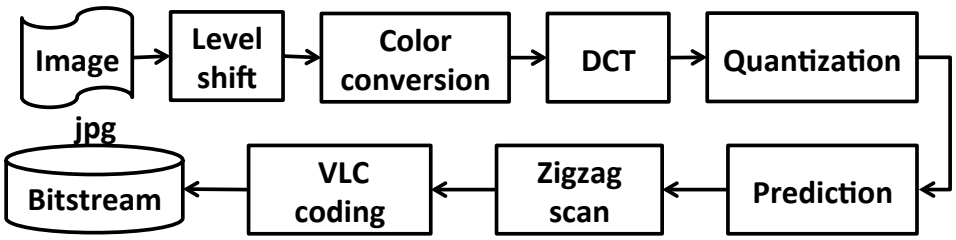


Figure 2.1: Process flow for JPEG compression (encoding phase)

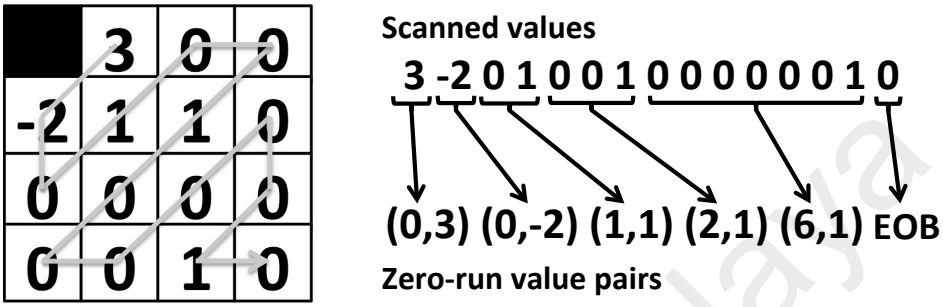


Figure 2.2: An example of forming ZRV using Zigzag scan (ISO/IEC, 1994)

resolution of the channels (i.e., Cb and Cr) of chrominance optionally can be reduced (sub-sampled) to achieve more compression efficiency. Next, non-overlapping blocks (i.e., 8×8 pixels) are obtained by partitioning each the (sub-sampled) channel to conduct block-based operations. Each the block is transformed by Discrete Cosine Transform (DCT) (K.R.Rao & P.Yip, 1990) to obtain coefficients in the frequency domain, where low-frequency coefficient values are generally perceptive to human eyes, and verse visa. Since high frequency coefficients are less perceptual to human eye, quantization operation, which utilizes a table referred as Quantization Table (QT) determined by the Quality Factor (QF), is applied to the obtained coefficients of each block for further removing redundancy of the coefficients. Then, the quantized Direct Current (DC) coefficient undergoes Differential Pulse Code Modulation (DPCM) based prediction while the Zigzag scan is applied to the quantized Alternating Current (AC) coefficients for forming Valuable Length Coding (VLC). Figure 2.2 illustrates an example of forming ZRV using the Zigzag scan. Finally, the outputs of them are entropy coded using VLC such as Huffman coding.

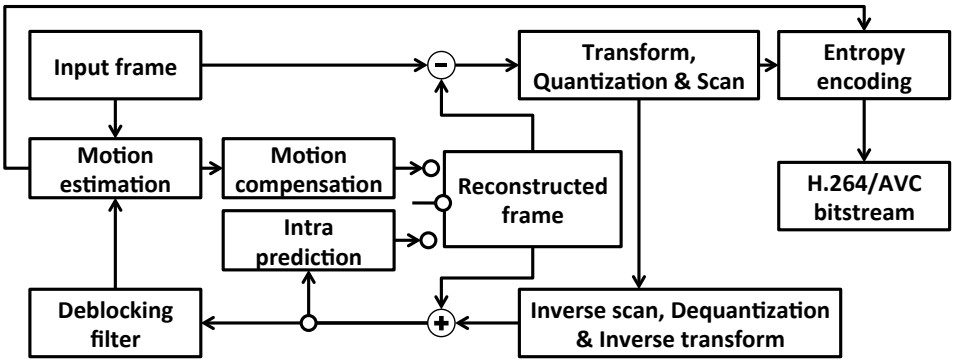


Figure 2.3: A flow of H.264/AVC encoder (Wiegand et al., 2003)

2.2.2 H.264/AVC

A general video encoder transforms a video formed with a sequence of frames into a compressed video bitstream. Figure 2.3 shows the flow of a general H.264/AVC encoder (Wiegand et al., 2003). H.264/AVC encoder selects an encoding frame mode (either I-, P- or B- frame mode) for each input frame according to the frame sequence. Next, The frames are partitioned into non-overlapping MBs, which are 16×16 pixel blocks. Then, each MB is encoded following its encoding frame mode. Specifically, I-frame MBs are encoded without any reference MBs, while P-frame MBs are encoded with or without reference (I- or P-frame preceding the current frame) MBs. In encoding B-frame MBs, the encoding process is similar to the process of P-frame MBs, but the encoding refers not only preceding I- or P-frame, but also I- or P-frames following the current frame.

More specifically, each MB in I-frame is further partitioned into blocks, which are either sixteen 4×4 pixel blocks, four 8×8, or one 16×16, are considered to refer its spatial complexity and then the partitioned MB information is recoded as MB Type (MBT). To reduce spatial redundancy, the pixel values of the MB are predicted from same frame neighbor blocks, namely North-East (NE), North (N), North-West (NW), and/or West (W) neighbors. Each direction information in used for prediction is recorded in INTRA Prediction Mode (INTRA-M). Then, the predicted values of each block are transformed with Integer Discrete Cosine Transform (IntDCT) and then quantized with QP. By using

either Context Adaptive Valuable Length Coding (CAVLC) or Context Adaptive Binary Arithmetic Coding (CABAC), the output of the aforementioned operations is entropy coded.

On the other hand, in encoding each MB of P- or B-frame, each MB is further partitioned into various size blocks, i.e., 4×4, 8×4, 8×8, 8×16, 16×8 or 16×16 according to the result of the quarter-pixel MV estimation is operated using single/multiple reference frames. Here, prior to motion estimation, in-loop Deblocking Filter (DF) is utilized for mitigating blocking noise, which is an intrinsic drawback of any block-transform compression approach. Next, the predicted values of each block are calculated, and entropy coded. Information of the output values and the information of the process modes are coded as H.264/AVC syntax.

H.264/AVC supports various Network Abstract Layer Unit (NALU) to carry the header information for transport layer or storage media as H.264/AVC syntax. These units include sequence parameter set that contains the input video parameters, picture parameter set that contains frame parameters, instantaneous decoder refresh that signals the reference frame buffer to be cleared, and several slices. As for slice level, each slice contains slice header and slice data, namely, end of sequence indicating the end of a frame and end of the stream indicating the end point of the sequence. Each H.264/AVC syntax can be placed into a logical data packet.

2.2.3 Symbols in the block-transform Domain

To avoid loss of generality in this thesis, an input g (an image or a frame of video) of dimension $X \times Y$ pixels is assumed. $g(x,y)$ indicates the (x,y)-th pixel value in the input g. In the compressed domain, generally, pre-processing (e.g., prediction process) is operated on the input g, and then the block transformed values $\mathbb{T}$ is obtained by using a block transformation with block size $\gamma \times \gamma$ pixels, where $\gamma = 4$ or 8. To gain compression, the

Table 2.1: Components Stored in the Bitstream of Compressed Image/Video

Standard	MBT	MV	PM	Coefficient	QC	EC	BPH
JPEG	×	×	×	✓	✓	✓	✓
H.264/AVC	✓	✓	✓	✓	✓	✓	✓

transformed values $\mathbb{T}$ are quantized. Hence, the (u,v)-th transformed value $\mathbb{T}_{u,v}(i,j)$ in the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$ can be denoted as follows:

$$\mathbb{T}_{u,v}(i,j) = constant \times \mathbb{F}_{u,v}(i,j) \times QT[u][v], \tag{2.1}$$

where $i \in [1,M]$, $j \in [1,N]$, $u,v \in [1,\gamma]$, $\gamma M = X$, $\gamma N = Y$, *constant* is a constant value, $\mathbb{F}_{u,v}(i,j)$ is the (u,v)-th transformed and quantized value in the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$, and $QT[u][v]$ is the (u,v)-th quantization table.

2.2.4 Components Stored in Compressed Image/Video

Bitstream of a compressed image/video contains various components / information, which can be categorized as follows: (A) MBT information, which indicates the size of MB for H.264/AVC; (B) MV information, which includes Differential Motion Vector (DMV) in H.264/AVC; (C) Prediction Mode (PM) mode information which includes INTRA-M and INTER Prediction Mode (INTER-M) in H.264/AVC; (D) Coefficient information, which includes the sign of DCT and its magnitude value; (E) Quality Control Parameter (QC) information, which includes QF in JPEG or QP, bitrate, and DF in H.264/AVC; (F) Entropy Code (EC), which includes VLC in JPEG, and CAVLC or CABAC in H.264/AVC, and; (G) Bitstream Packet Header (BPH) information, which includes group of picture, resolution, and NALU information. Table 2.1 summarizes the aforementioned components/information encoded within a compressed image/video bitstream. Here, if the box of row α and column β is marked with ✓, it signifies that standard α includes component β in compressed bitstream, and × signifies otherwise.

2.3 Overview of Encryption

As the demand for high-resolution images and HD videos increases, practical security tools also become significantly important for JPEG compressed image and H.264/AVC compressed video. In image/video compressed content, perceptual information quality is one of significant property. To handle the property in perceptual encryption, Thomas et al. (Stutz & Uhl, 2012) formalized transparent encryption, which is able to control the perceptual quality of a video. In addition, this perceptual quality control ability is practical for applications, such as VOD, VSS, VC, VA, SS. The reason is that desired distortion videos can be flexibly generated by the content provider without revealing the high-quality content before transmission to potential purchasers.

Along with the emergence of advanced compression techniques as well as their applications, the rate at which compressed images/videos are generated is increased, and the demand security tool is also increased. Applying a classic cipher, i.e., Data Encryption Standard (DES), Advanced Encryption Standard (AES) or Rivest-Shamir-Adleman (RSA), to the bitstream of a compressed image/video is an intuitive security tool. However, this class of approaches is too complex and fails various applications requirements. To overcome the classic cipher problems, a number of image/video encryption algorithms and their classifications (Socek et al., 2007; Massoudi et al., 2008; Stutz & Uhl, 2012; Padilla-López et al., 2015) have been proposed. The interpretation and classification might vary from one author to another. To avoid misinterpretation, three classifications are defined in this thesis.

The first class is the term of selected target for encryption: Full encryption (T_1), which operates on the entire bitstream, and; Selective encryption (T_2), which operates on selected bits of the bitstream. Next, the second class necessitates the algorithm to be format-compliance (Jiangtao et al., 2002) (transparency (Pazarci & Dipcin, 2002),

or syntax-awareness (H. Li & Jian Ren, 2007)). Here, Format-compliant (F_1) refers to the case where the general decoder can decode the encrypted images/videos, and; Non format-compliant (F_2) refers that a general decoder can't decode the encrypted images/videos. The third class is in terms of encryption domain: Uncompressed domain (D_1), which means that encryption is operated before compression process; Compressed domain (D_2), which means that encryption is operated jointly with compression process, and; Bitstream domain (D_3), which means that encryption handles encoded bitstream directly.

Most of the conventional encryption methods (Massoudi et al., 2008; Stutz & Uhl, 2012) meet to both of Class T_1 and Class F_1 in the literature. I refer them as the term *format-compliant selective encryption* in this thesis. In addition, in Class D_1, the image/video is encrypted before compression processing. However, it is intuitively expected that the class compromises the compression performance. The reason is that encryption introduces randomness, which makes difficult to exploit redundancy by compression operations. Furthermore, Class D_2 encrypts the components, such as sign information, integer transformed coefficients and scanning order, in the image/video compression standard. Moreover, Class D_3 encrypts the encoded bitstream, but this class can encrypt only limited bits, e.g., packet body in Joint Picture Expert Group 2000 (JPEG2000) and regular units of network abstraction layer units in H.264/AVC. Therefore, Class D_2 is the promising approach to achieve the aforementioned advanced application scenarios. This study focuses on format-compliant selective encryption of Class D_2. The diagram of the aforementioned classification can be illustrated in Figure 2.4.

2.3.1 Format-Compliant Selective Encryption for JPEG

In the literature, seven elementary encryption modules, which handle the components in JPEG, have been considered to design format-compliant selective encryption. Con-

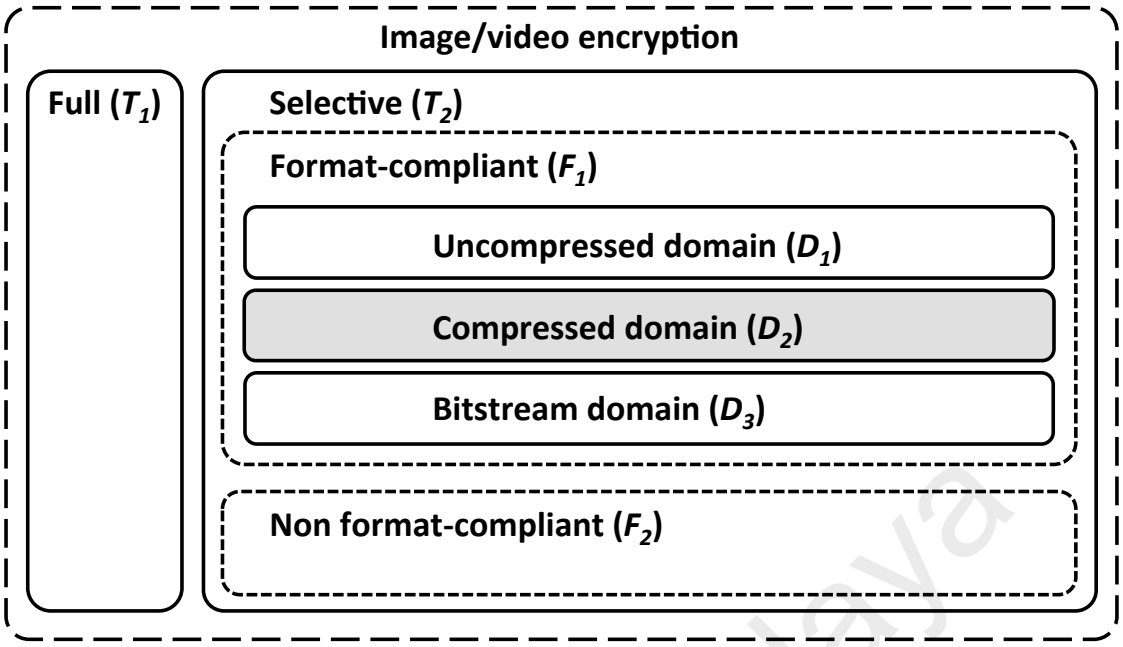


Figure 2.4: Classification of image/video encryption classes (Socek et al., 2007; Massoudi et al., 2008; Stutz & Uhl, 2012; Padilla-López et al., 2015)

ventional methods selectively considered these elementary modules. In the following sub-sections, each the module is briefly described.

2.3.1.1 DC Shuffling (DCS)

DCS (W. Zeng & Lei, 2003) shuffles DC coefficients within the window of a certain size. The module is able to control the outline quality of the distorted image, but it causes bitstream size overhead.

2.3.1.2 DC Error Mapping (DCEM)

DCEM (Niu et al., 2008) maps each DC residue value processed by DPCM to a different value, which belongs to the same code category. Although this module can maintain the bitstream size, it cannot withstand the sketch attack (W. Li & Yuan, 2007).

2.3.1.3 DC Sign Randomization (DCSR)

DCSR (Shiguo et al., 2004) randomly flips the sign of DC coefficients.

2.3.1.4 AC Sign Randomization (ACSR)

Each sign of AC coefficient is randomly flipped by ACSR (Lian et al., 2007). Since the sign information is coded as one bit, ACSR can maintain the bitstream size. However, the output of ACSR fails to withstand the sketch attack (W. Li & Yuan, 2007).

2.3.1.5 AC Subband Shuffling (ACSS)

The sequence of the AC coefficients is rearranged by ACSS $\mathbb{F}'_{u,v}(i,j)$ (W. Li & Yuan, 2007; Shiguo et al., 2004). Let $\mathbb{F}'_{u,v}(i,j)$ denote the shuffled (u,v)-th transformed coefficient of the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$. This module conducted within in the same frequency band can be expressed as Equation (2.2):

$$\mathbb{F}'_{u,v}(i,j) = \mathbb{F}_{u,v}(i',j'), \tag{2.2}$$

and this module conducted within the same block can be expressed as Equation (2.3):

$$\mathbb{F}'_{u,v}(i,j) = \mathbb{F}_{u',v'}(i,j), \tag{2.3}$$

where each symbol is defined as follows: $1 \leq i,i' \leq M$; $1 \leq j,j' \leq N$, and; $1 \leq u,v,u',v' \leq 8$. Although, by shuffling with Equation (2.2) and Equation (2.3), sufficient-distortion images can be obtained, the bitstream size of the output severely increases (e.g., $+\sim 20\%$ in Reference (W. Li & Yuan, 2007). Note that Equation (2.2) can withstand sketch attacks, while Equation (2.3) fails.

2.3.1.6 AC ZRV Pair Scrambling (ACPS)

ACPS (Takayama et al., 2006; Wong & Tanaka, 2010) shuffles the order of Zero-Run Value (ZRV) pairs, and there are two approaches: INTRA-block shuffling, and; INTER-block shuffling. In INTER-block shuffling approach (Wong & Tanaka, 2010), all ZRV pairs are extracted from the image, and then they are shuffled and distributed to each

block. Although INTER-block shuffling approach can achieve sufficient distortion for encryption, extra information, (namely, the number of nonzero coefficients of each block) is required for the restoration operations. On the other hand, in INTRA-block shuffling approach (Takayama et al., 2006), all ZRV pairs within each block are shuffled. This approach doesn't require extra information but is vulnerable to sketch attack.

2.3.1.7 AC Block Shuffling (ACBS)

ACBS (Takayama et al., 2006) permutes AC coefficient blocks within a window of a certain size. Let $\mathbb{F}_{AC}(i,j)$ denote the (i,j)-th AC block, i.e., the (i,j)-th block $\mathbb{F}(i,j)$ in $\mathbb{T}$ without the DC coefficient. ACBS can be expressed by Equation (2.5):

$$\mathbb{F}^{'}_{AC}(i,j) = \mathbb{F}_{AC}(i',j'), \tag{2.4}$$

This approach sufficiently distorts the input image and withstand against the sketch attack (W. Li & Yuan, 2007), because the edge information is shuffled on the entire image. In addition, this approach causes insignificant bitstream size overhead due to byte-alignment (B.Pennebaker & L.Mitchell, 1992). However, ACBS is insufficient because the outline image of the original image can be revealed from the intact DC coefficients.

2.3.1.8 Block Shuffling (BS)

BS (Niu et al., 2008) permutes DCT coefficient blocks as expressed by Equation (2.5) within a window of a certain size:

$$\mathbb{F}^{'}(i,j) = \mathbb{F}(i',j'), \tag{2.5}$$

where $\mathbb{F}^{'}(i,j)$ denotes the shuffled (i,j)-th block in $\mathbb{T}$. Since the order of blocks is shuffled to transform texture/edge information to unintelligible form, BS can sufficiently distort the image quality and withstand sketch attack (W. Li & Yuan, 2007). Note that, this

module changes bitstream size due to byte-alignment (B.Pennebaker & L.Mitchell, 1992).

2.3.1.9 Coefficient Value Mapping (CVM)

CVM alters a DCT coefficient value regardless nonzero and zero coefficient value. Reference (M. Zhang & Tong, 2014) put forward an encryption method using a chaotic map, which alters all partially decoded DCT coefficient values of each color channel to different values. In other words, this approach maps all DC and AC coefficient values to different values. Similarly, Takayama et al. (Takayama et al., 2008) proposed a bijectively mapping method that maps each AC coefficient to a different value. This approach uses a parameter to generate a bijective map. Although these approaches achieve significant distortion by changing the magnitude of the coefficient values, the process causes severe bitstream size overhead.

The aforementioned modules are selectively included in the conventional encryption methods, which are summarized in Table 2.2. Note that, when the row α and the column β is marked with ✓, the method at the row α includes the module at the column β. On the other hand, when α and the column β is marked with $\times$, the method at the row α doesn't include the module at the column β.

Table 2.2: Format-compliant selective encryption modules for JPEG image

	DCS	DCEM	DCSR	ACSR	ACSS	ACPS	ACBS	BS	CVM
Van Droogenbroeck & Benedett (2002)	×	×	×	✓	×	×	×	×	×
(video) W. Zeng & Lei (2003)	✓	×	✓	✓	×	×	×	×	×
Shiguo et al. (2004)	×	×	×	✓	✓	×	×	×	×
(video) Takayama et al. (2006)	✓	×	×	×	×	✓	✓	×	×
W. Li & Yuan (2007)	×	×	×	×	✓	×	×	×	×
(video) Lian et al. (2007)	×	×	✓	✓	×	×	×	×	×
Niu et al. (2008)	×	✓	×	×	×	×	×	✓	×
(video) Takayama et al. (2008)	×	×	×	×	×	✓	✓	×	✓
Wong & Tanaka (2010)	✓	×	×	×	×	✓	×	×	×
M. Zhang & Tong (2014)	×	×	×	×	×	×	×	×	✓

2.3.2 Format-Compliant Selective Encryption for H.264/AVC video

In H.264/AVC compressed video encryption, nine elementary modules handling the compressed video components are considered in the literature (Stutz & Uhl, 2012). These elementary modules were selectively considered to design desired effects for conventional format-compliant selective encryption methods. In the following sub-sections, each elementary module is briefly described.

2.3.2.1 IntDCT Coefficient Sign Randomization (S)

S (Shiguo et al., 2006) operates sign flipping, which modifies the sign (i.e., plus or minus) of each coefficient with a pseudo-random sequence. This module maintains the bitstream size of the output encoded with CAVLC.

2.3.2.2 IntDCT Coefficient Level Modification (L)

L (Magli et al., 2006) maps the nonzero coefficient values to different coefficient values before entropy coding operation. Specifically, this module modifies the magnitude of the coefficient values.

2.3.2.3 Secret Transformation (ST)

ST (Siu et al., 2009) randomly selects a 4×4 integer transformation function from either IntDCT or an other integer transformation, such as integer discrete sign transformation.

2.3.2.4 Quantization Parameter Manipulation (QPM)

QPM (Spinsante et al., 2005) modifies the quantization parameters of each MB to alternate the reconstructed coefficient values.

2.3.2.5 Deblocking Filter Value Modification (DFV)

H.264/AVC encoder utilizes the deblocking filter to reduce block noise before the motion estimation operation and the compensation operation. DFV (Siu et al., 2009) is designed

to alter the global deblocking filter value for achieving distortion effect.

2.3.2.6 MV Difference Modification (MVD)

MVD (Su et al., 2011) modifies the motion vectors and their differences.

2.3.2.7 Secret Scan Order (SSO)

To encode coefficient values in each MB, the 4×4 Zigzag scan is utilized in H.264/AVC. SSO (Su et al., 2011) utilizes difference scan orders to achieve distortion effect.

2.3.2.8 INTER-Prediction Mode Modification (INTER-MM)

In H.264/AVC, various the subdivisions/Sub MB (subMB), namely number of 16×16, 16×8, 8×16 and 8×8 blocks, are utilized to facilitate compression gain. INTER-MM (Yuan et al., 2005) proposed a shuffling the prediction mode of the compatible structure MBs in INTER-frame, because simply shuffling difference structure MB causes decoding problems.

2.3.2.9 INTRA-Prediction Mode Modification (INTRA-MM)

Similarly to INTER-MM, by modifying intra prediction mode of the input video, distorted videos can be obtained. INTRA-MM (Yuan et al., 2005) proposed a prediction mode modification prior to the transformation process (i.e., IntDCT). The aforementioned modules are selectively included in the conventional encryption methods, which are summarized in Table 2.3. Note that the interpretation of the Table is the same manner as in Table. 2.2. The recently proposed format-compliant selective encryption methods (Yongsheng et al., 2013; B. Zeng et al., 2014) claim that their methods are secure against cryptanalytic attacks.

Table 2.3: Format-compliant selective encryption modules for H.264/AVC video

	S	L	ST	QPM	DFV	MVD	SSO	INTER-MM	INTRA-MM
Lian et al. (2007)	✓	×	×	×	×	✓	×	×	✓
Thomas et al. (2007)	×	×	×	×	×	✓	×	×	×
Ahn et al. (2004)	×	×	×	×	×	×	×	×	✓
Yuan et al. (2005)	×	×	×	×	×	✓	×	✓	✓
Bergeron & Lamy-Bergot (2005)	×	×	×	×	×	×	×	×	✓
Shiguo et al. (2005)	✓	×	×	×	×	✓	×	×	✓
Shiguo et al. (2006)	✓	✓	×	×	×	✓	×	×	✓
Shiguo et al. (2008)	✓	×	×	×	×	✓	×	×	✓
Shahid et al. (2009)	✓	✓	×	×	×	×	×	×	✓
Su et al. (2011)	×	×	×	×	×	✓	✓	×	✓
Kwon et al. (2005)	×	×	×	×	×	✓	×	×	×
Magli et al. (2006)	×	✓	×	×	×	✓	×	×	×
Won et al. (2006)	✓	×	×	×	×	✓	×	×	×
Grangetto et al. (2007)	×	✓	×	×	×	✓	×	×	×
Yeongyun et al. (2007)	✓	×	×	×	×	✓	×	×	×
Yang et al. (2007)	×	×	×	×	×	✓	×	×	×
Siu et al. (2009)	×	×	✓	×	×	×	×	×	×
Spinsante et al. (2005)	×	×	×	✓	✓	×	×	×	✓
Lee & Nam (2006)	✓	×	×	×	×	×	×	×	×
Chunhua et al. (2008)	✓	×	×	×	×	×	×	×	×
Y. Wang et al. (2013)	✓	×	×	×	×	×	×	×	✓
Shen et al. (2014)	✓	×	×	×	×	✓	×	×	✓
Dubois et al. (2014)	✓	✓	×	×	×	×	×	×	×
Boho et al. (2013)	✓	×	×	×	×	✓	×	✓	✓
Peng et al. (2013)	✓	×	×	×	×	✓	×	×	✓
Yongsheng et al. (2013)	✓	×	×	×	×	✓	×	×	✓
B. Zeng et al. (2014)	×	×	✓	×	×	×	×	×	×

2.3.3 Cryptanalysis

Cryptographic attacks aim to reconstruct the original image/video from its encrypted image/video. Note that known-plaintext attack and chosen-plaintext attack are well known as cryptographic attacks. Thus, most conventional format-compliant selective encryption methods were designed to avoid information leaking by the cryptographic attacks. However, other attack approaches may success to obtain information of the original content. To my best knowledge, two perceptual attack approaches exist as follows: (A) error-concealment based attack, and (B) replacement attack. Approach (A) utilizes error-concealment techniques, which handle the encrypted parts as missing parts. Approach (B) substitutes the encrypted parts with attacker-defined data. These perceptual attack approaches can reconstruct only low-quality outputs of the original image/video from image/video encrypted by shuffling DCT coefficients.

Although the aforementioned format-compliant selective encryption methods satisfy conventional cryptanalysis and perceptual attack approaches, a perceptual attack (hereinafter referred to as sketch attack), Nonzero Coefficient Count Attack (NZCA) (W. Li & Yuan, 2007), is put forward to extract a set of rough contour lines of the original image directly from its JPEG encrypted images. Therefore, the sketch image by the sketch attack can show visual information extracted from encrypted images.

2.3.4 Conventional Sketch Attack

Although a novel attacking approach, namely sketch attack, is invented, the community put attention only to cryptographic attacks (Takayama et al., 2006; Wong & Tanaka, 2010; M. Zhang & Tong, 2014). In the surveyed literature for this thesis, there is only one sketch attack (NZCA) (W. Li & Yuan, 2007) in the compressed domain. NZCA (W. Li & Yuan, 2007) utilizes the number of nonzero coefficients in each 8×8 block to generate a sketch image ϕ_Z, which consists of the pseudo-luminance values calculated as follows:

$$\phi_Z(x,y) = \phi_Z^{'}(\lfloor i/8 \rfloor, \lfloor j/8 \rfloor) \tag{2.6}$$

$$\phi_Z^{'}(i,j) = \begin{cases} 0 & n(i,j) < n_1 \\ 255 & n_1 \leq n(i,j) \leq n_2 \end{cases} \tag{2.7}$$

where $\phi_Z^{'}(i,j)$ is the pseudo-luminance value in $\mathbb{F}(i,j)$, and $n(i,j)$ is the number of non-zero AC coefficients in the (i,j)-th 8×8 block in $\mathbb{F}(i,j)$, where the number of AC coefficients implies the spatial activity of a block. The two values, viz. n_1 and n_2, are the two threshold values to tune the sketch image, where proper threshold values will separate backgrounds and foregrounds. The sketch attack considers only AC coefficients attained by DCT, hence other components in the compressed domain, which can be used for feature extraction, still remained.

2.4 Overview of Outline Clearness Assessment

Although edge images (obtained from edge detector) and sketch images (obtained from sketch attack) are generated in different domains, they both show the outline of the image/video frame of interest. Therefore in this thesis, these images and their detectors are referred by using the terms *outline image* and *outline detector*, respectively.

A general process flow for outline-based applications can be summarized in Figure 2.5, where *input* indicates image/video frame data and *output* indicates the results of the applications. This flow is utilized in advanced automated applications, including textual enhancement (Chi & Eramian, 2015), text detection (Akhaee et al., 2010; Shivakumara et al., 2011), and image segmentation (Tong et al., 2005). However, outline-based

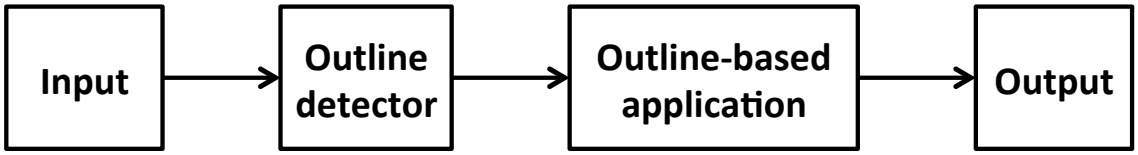


Figure 2.5: General process flow of outline-based application system

application systems may not be robust against the intrinsic noise generated by outline detectors, since outline-based applications are not designed to process noisy outline images. Therefore, a reliable OCA is crucial in ensuring proper operability of the outline-based application system of interest. Specifically, the goal of OCA is to objectively evaluate the clearness of an outline image.

An obvious way of measuring outline clearness is to consider a subjective evaluation by humans. However, such subjective evaluations are not only slow, cumbersome, time-consuming and expensive, but it is also incapable for deployment in real-time applications that adaptively adjust the operational parameters for improving application performance. Some learning-based outline detectors are trained to detect outline from the input image, but only clear images are presented during the training stage. Therefore, when the input image is noisy, these outline detectors may not perform in the intended manner. Furthermore, since the ideal outline cannot be obtained in practical situations, it is futile to have a reference based OCA metric.

In the literature, there are two classes of domain outline detectors, i.e., edge detector and sketch attack, In addition, there are also two types of assessments, i.e., IQA for natural images and an edge similarity assessment (W. Li & Yuan, 2007) for binary sketch images.

2.4.1 Overview of Outline Detector

Traditionally, the edge map is deployed in image processing applications, while the sketched image is utilized to attack compressed-then-encrypted image (see Section 2.3.4). In this section, the conventional edge detectors are briefly summarized.

(a) Class (A) - CAN (b) Class(B) - SFE

Figure 2.6: Outline generated by methods in Class (A) and (B)

Edge refers to the collection of points where pixel intensities change abruptly. In the surveyed literature for this thesis, edge detectors can be grouped into two categories (Y. Li et al., 2015), viz.: (A) differentiation based, and; (B) learning based. Class (A) considers the discontinuities of brightness in an image, while Class (B) uses machine learning approach to discriminate edge points from smooth regions by considering multiple low-level cues extracted from the training images. In contrast to metrics in Class (A), metrics in Class (B) is viable to alleviate the internal edges on textures. It is observed that the internal edges of Figure 2.6(a) are removed in Figure 2.6(b). In addition, the edge detections in Class (B) are designed to resemble the human visual system perception, thus the output image has a more semantic meaning of the image (Y. Li et al., 2015).

Class (A) edge detectors are designed to capture gradient in the input image. Specifically, Sobel Edge Detector (SOB) (Gonzalez & Woods, 2006) is designed to find local maxima of the first-order gradient at different orientations. Laplacian of Gaussian (LOG) (Gonzalez & Woods, 2006) is designed to find zero-crossing points of second-order derivative, where Gaussian filter is applied for smoothing the input image because derivative operations are sensitive to noise. CAN (Canny, 1986) is designed to find the

solution to an optimization problem with three criteria, viz., single-pixel response, localization, detection accuracy. In addition, CAN outperforms several detectors, including a recent metric (Y. Li et al., 2015), and it is still widely deployed in various applications. On the other hand, Structure Forest Based Edge Detector (SFE) (Dollar & Zitnick, 2013) from Class (B) learns color, gradient, and pair-wise differences by using structured random forests derived from random decision forests (Geurts et al., 2006). By using the learning results, the edge pixels of the input patch are allocated to structure random forests. The edge responses are finally obtained as an aggregation (in this case averaging) of the random forests. SFE is relatively low in computational complexity while achieving comparable accuracy to the currently proposed methods.

2.4.2 IQA

Following the literature of IQA, in terms of reference dependency, all assessment methods can be grouped into three classes (Lin & Jay Kuo, 2011), viz.: (A) full-reference based; (B) reduced-reference based, and; (C) no-reference based. Class (A) utilizes the features extracted to the reference (original) image to assess the processed images. Class (B) utilizes the features of some part(s) in the reference image. On the other hand, no information about the reference image is required by Class (C), thus the fact make Class (C) is the most attractive assessment metric in the image processing research community.

Various no-reference IQA methods (Attar et al., 2015; Guan et al., 2015; L. Li et al., 2015; F. Qian et al., 2015; Q. Wu et al., 2015) are proposed recently, and many of them achieve promising results when handling natural images. However, the features considered in these no-reference IQA methods may not be directly applicable to evaluate outline image (further discussed in Section 4.3). These shortcomings of the conventional no-reference IQA methods motivated us to form a no-reference assessment method for outline image quality evaluation.

2.4.3 Edge Similarity

In addition to the sketch attack (NZCA), Li et al. (W. Li & Yuan, 2007) proposed an edge similarity evaluation method, which considers the SOB edge map Γ_0 (Gonzalez & Woods, 2006) as the reference edge map. Note that, since the NZCA sketch image is binary form, Otsu thresholding (Otsu, 1979) is applied to the SOB edge map to generate a binary image, and then the output is resized because the sketch image is resolution smaller than the binary SOB edge map. Let ζ_0 denote the value obtained by the evaluation method. The edge similarity equation can be expressed as follows:

$$\zeta_0 = \frac{\eta_1}{2\Delta_0} + \frac{\eta_2}{2\Theta_0} - \frac{\eta_3}{2\Delta_0} - \frac{\eta_4}{2\Theta_0}, \tag{2.8}$$

where Δ_0 denotes the total number of 0's in SOB edge map Γ_0 and Θ_0 denotes the total number of 1's. In addition, the definitions of η_1, η_2, η_3 and η_4 are as follows, respectively:

$$\eta_1 = |\{(x,y) : \Gamma_0(x,y) = 0 \ \& \ \phi(x,y) = 0\}|, \tag{2.9}$$

$$\eta_2 = |\{(x,y) : \Gamma_0(x,y) = 1 \ \& \ \phi(x,y) = 1\}|, \tag{2.10}$$

$$\eta_3 = |\{(x,y) : \Gamma_0(x,y) = 0 \ \& \ \phi(x,y) = 1\}|, \tag{2.11}$$

$$\eta_4 = |\{(x,y) : \Gamma_0(x,y) = 1 \ \& \ \phi(x,y) = 0\}|. \tag{2.12}$$

where ϕ denotes a sketch image. The particular behaviors of Li et al.'s (W. Li & Yuan, 2007) edge similarity evaluation method can be described as follows: (a) $\zeta_0 = 1$ indicates that the sketch image and the reference edge map, in Li et al (W. Li & Yuan, 2007) case, the scaled SOB edge map, are exactly the same; (b) $\zeta_0 = -1$ indicates that the sketch image is exactly the negated map of the reference edge map, and; (c) $\zeta_0 = 0$ indicates that

either the sketch image is random or the sketch image is completely black (0's) or white (i.e., 1's).

2.5 Overview of Text Detection in Video

Text contained in video provides related information which enriches the video contents, hence the text information can be utilized to video applications, i.e., video indexing and retrieval (J. Zhang & Kasturi, 2008; Jung et al., 2004). In addition, the text information in video can be utilized as features for labeling the events of the video contents, and then the performance of the text information based methods are better when comparing them with content based video retrieval methods based on semantics (J. Zhang & Kasturi, 2008; Jung et al., 2004). Therefore, text detection is arguably the most crucial step for the aforementioned applications, (i.e., video indexing and retrieval).

Text in video can be classified into two categories, namely (A) graphics text , and; (B) scene text. Graphics text is intentionally inserted text, i.e., subtitle and/or during video production. In contrast, scene text is unintentionally captured text as a part of a natural scene. Generally, graphics text due to text insertion is perceptually clear, in other words, the video is high contrast. On the other hand, since scene text is captured from a natural scene, scene text appears in various forms, such as font size variations, font type, background, contrast, orientation, etc. Hence, text detection in scene text is a challenging task. In particularly, multi-oriented text detection is technically challenging because estimating the orientation of the text is difficult and is not readily available. Therefore, the research community is still pursuing multi-oriented text detection.

In the literature, text detection domain can be classified into the non-compressed domain and the compressed domain. The following sub-sections review the conventional methods in both domains.

2.5.1 Overview of Text Detection in the Spatial Domain

In this sub-section, the representative conventional text detection (including localization) methods in the spatial domain are overviewed. Text detection in the spatial domain can be classified into three categories: (a) component based methods (Jain & Yu, 1998); (b) texture based methods (Shivakumara et al., 2010, 2011), and; (c) edge and gradient based methods (C. Liu et al., 2005; Shivakumara et al., 2012, 2013). First, component-based methods perform well for caption and graphics text when these texts have uniform color and high contrast. Next, texture based methods achieved sufficient detection performances for complex background, however, these performances highly depend on the font size and font type of the text as well as the computational cost is high. Last, edge and gradient based methods achieved less computational cost against to texture based methods. However, these methods are sensitive to cluttered background and thus produce more false positives. In the aforementioned text detection methods, horizontal graphics text has been focused while multi-oriented scene text has been put less attention.

Recently, few methods (Shivakumara et al., 2012, 2013) considering both graphics and scene multi-oriented text have been proposed. However, the major issue of these methods is that their accuracies are inconsistent across different image datasets. In addition, the text detection methods are designed by considering only the pixel values in spite of the fact that most videos including text are stored in the compressed forms, which consist of compressed components (e.g., coefficients). Since these features (such as transformed coefficients, motion vectors and MBs type) have information, the components in the compressed domain are practical for text detection. Furthermore, by using components partially decoded from the compressed domain (Zhong et al., 1999), reduction of the computational cost and storage space can be expected.

2.5.2 Overview of Text Detection in the Compressed Domain

Comparing to the number of text detection methods in the spatial domain have been investigated, the number of text detection in the compressed domain is small. In this subsection, the representative text detection methods in the Moving Picture Experts Group (MPEG) domain are reviewed as following the published order. Early stage automated text detection methods had focused on detecting caption text (i.e., subtitle or short description text). First, the text detection proposed by Zhong et al. (Zhong et al., 1999) utilizes the information of entities in the MPEG domain and the results of a connected component analysis, however, they put less attention to oriented scene text. Next, Dimitrova et al. (Dimitrova et al., 2000) considered the edge information induced from the DCT coefficients and proposed a superimposed text detection in video. This method also put attention to graphics text but not scene text in a video. Then, to achieve text detection and tracking in the compressed domain, the information of text appearance and disappearance in video frames is considered by Qian et al. (X. Qian et al., 2007). It is observed that this approach is viable to detect horizontal graphics text. Then, text detection and tracking are further improved by Jiang et al. (Jiang et al., 2008) to achieve a faster method. Since their method considers the defined starting and ending frames, and utilizes an initial threshold value for text matching, the performance depends on the threshold value. Last, the DCT-based features are investigated by Goto et al. (Goto, 2008) to design scene text detection, however, their method is only viable to horizontal text with high contrast.

Although, some text detection methods in the MPEG compressed domain are proposed, the current most deployed video coding standard is H.264/AVC. A text detection in H.264/AVC compressed video is proposed by Qian et al. (X. Qian et al., 2012) in the surveyed literature. Although this method is designed to detect text in INTRA-frame, it depends on threshold values for text block verification and it also targets on graphics text

of horizontal direction only.

2.6 Overview of Sketch Attack

The surveyed conventional format-compliant selective encryption methods may be vulnerable to sketch attack (W. Li & Yuan, 2007), because some of the components stored in the compressed domain remained intact. This fact suggests that all known sketch attacks must be investigated and then a countermeasure should also be put forward to prevent perceptual information leakage.

In addition, edge images play a significant role in multimedia applications, viz., enhancement (Chi & Eramian, 2015), segmentation (Tong et al., 2005), text detection (Shivakumara et al., 2011), digital watermarking (Akhaee et al., 2010) and image retrieval (Q. Li et al., 2016). Similarly, sketch images by sketch attacks may also play a significant role in the aforementioned multimedia applications. Therefore, an assessment criterion of the outline image, i.e., edge image and sketch image, is crucial to multimedia applications.

Furthermore, multimedia applications, including text detection (X. Qian et al., 2007), encryption, visible watermarking, data hiding and foreground-background separation (Dey & Kundu, 2013) in the compressed domain have been actively pursued by researchers recently. As a new feature extraction process in the compressed domain, it is worth exploring sketch attack in related applications. Specifically, the study of new sketch attacks, the study of outline assessment and the demonstration of sketch image based application will be pursued.

2.7 Summary

The widely-deployed compression standards, including JPEG and H.264/AVC, are briefly described and their format-compliant selective encryption and cryptanalysis are surveyed. A novel perceptual attack, namely sketch attack, is introduced and the weakness of conventional format-compliant selective encryption is analyzed. The analysis indicates that,

after encryption, conventional encryption methods untouched some components. In other words, the outline information of the original image can be extracted. On the other hand, there are many multimedia applications based on edge/sketched images and the demand of signal processing in the compressed domain. These facts justify the need for this study.

CHAPTER 3: SKETCH ATTACKS FOR FORMAT-COMPLIANT SELECTIVELY ENCRYPTED VIDEO

3.1 Overview

In this chapter, five novel sketch attacks are proposed to attack H.264/AVC videos encrypted by conventional format-compliant selective encryption methods. First, the limitations of coefficient-based sketch attacks for H.264/AVC video in encrypted form are described. In particularly, the performance of sketching outline image of the coefficient-based sketch attacks is significantly low when INTER-frame is considered, where INTER-frame includes significantly reduced temporal information. Next, a sketch attack considering the number of bits spent on coding an MB is put forward to sketch outline image from INTER-frame. Then, the Canny edge map is considered as an ideal (reference) outline image to evaluate the sketch image. Last, using some video datasets, experiments are conducted to verify the performance of the five proposed sketch attacks. Results suggest that the sketch image of the MB bit based sketch attack can reveal an outline of not only INTRA-frame but also INTER-frame.

3.2 Introduction

As mentioned in Section 2.3.2, generally conventional format-compliant selective video encryption methods meet the security requirements against cryptographic attacks, such as known/chosen-plaintext attacks. In addition, these encryption methods can be improved by shuffling DCT coefficients within a block to withstand two kinds of perceptual attacks: (a) error-concealment based attack (Jiangtao et al., 2002), and; (b) replacement attack (C.-P. Wu & Kuo, 2005; Podesser et al., 2002).

However, to generate an outline image of the original (i.e., plaintext) JPEG image directly from its encrypted counterpart, a sketch attack (i.e., NZCA) is proposed by Li et

at. (W. Li & Yuan, 2007). Therefore, in terms of sketch attack, format-compliant selective video encryption may not be completely secure.

In the following sections, 5 sketch attacks, including: DC Error Category Attack (DCEC); Improved Nonzero Coefficient Count Attack (INCC); Position of the Last Nonzero Coefficient Attack (PLZ); Sum of Absolute AC Coefficients Attack (SAC), and; MB Bitstream Size Attack (MBS), are put forward to generate outline images directly from H.264/AVC compressed-only or compressed then format-compliant selectively encrypted (hereinafter, compressed-encrypted) videos. Then, the sketch images are evaluated by using the modified Edge Similarity Score (ESS) and the performances are discussed.

3.3 Proposed Sketch Attacks

Based on the literature survey carried out in this study, the method proposed in (W. Li & Yuan, 2007) is the only sketch attack in the literature. The sketch attack can reveal an outline image of the original image, however, the sketch attack requires manual tuning of two threshold values, which limits the sketch attack to a manual operation. In addition, the DC coefficients and the bits allocated to each MB have not been put attention by researchers. Therefore, first, the DCEC, is put forward to verify the robustness of DC-based format-compliant selective encryption method against unauthorized viewing. Next, three threshold-free AC-based sketch attacks, namely, INCC, PLZ and SAC, are put forward. Finally, MBS, is put forward. In this chapter, H.264/AVC compressed video in level 5.1 and Baseline Profile (BP) is considered to facilitate the discussion and avoid loss of generality. Specifically, BP utilizes only one slice within a frame, 4×4 IntDCT and 7 block sizes, viz., 4×4, 4×8, 8×4, 8×8, 8×16, 16×8 and 16×16 for H.264/AVC video compression. The novel five sketch attacks are detailed to generate the outline images of H.264/AVC compressed video in the following sub-sections.

3.3.1 DC Error Category Attack (DCEC)

DC coefficient value is the mean value of a transform block, hence, the DC coefficient array of a video frame is essentially the original video frame. However, the resolution of the array is smaller (i.e., $1/4$ in the case of H.264/AVC) than that of the original frame. In H.264/AVC video compression, the DC coefficients are processed by the DC prediction function similar to that of edge detection. Thus, some edge information is carried by the prediction errors of DC coefficients. Depending on their magnitudes, the prediction errors are grouped into categories for coding purposes. Therefore, by considering a representative value for each category, a sketch image ϕ_D can be generated as follows:

$$\phi_D(i,j) = 2^{d(i,j)}, \tag{3.1}$$

$$d(i,j) = \begin{cases} 2^{\log_2 |r(i,j)|} & \text{if } |r(i,j)| > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where (i,j) is the coordinate of the (i,j)-th 4×4 block and $d(i,j)$ is the category of the residue DC value $r(i,j)$ (i.e., prediction error).

3.3.2 Improved Nonzero Coefficient Count Attack (INCC)

NZCA utilizes each the number of nonzero coefficients in each block, but this attack requires manual thresholding. To solve this problem, a sketch image ϕ_N by INCC is proposed as follows:

$$\phi_N(i,j) = \text{round}\left(255 \times \frac{n(i,j)}{\max\{n(i,j)\}}\right), \tag{3.3}$$

where the number of nonzero AC coefficients in the (i,j)-th 4×4 block denotes as $n(i,j)$.

The number $n(i,j)$ infers the complexity of the (i,j)-th 4×4 block, in other words, a small $n(i,j)$ implies low spatial activity within the block, and vice versa.

3.3.3 Position of The Last Nonzero Coefficient Attack (PLZ)

PLZ generates a sketch image ϕ_P by the position of the last nonzero coefficient of each block, where, the Zigzag order is considered. The equation is as follows:

$$\phi_P(i,j) = \text{round}\left(255 \times \frac{p(i,j)}{\max\{p(i,j)\}}\right), \tag{3.4}$$

where the position of the last nonzero AC coefficients in the (i,j)-th block is denoted as $p(i,j)$. The value behaves as follows: (a) a large value when the block has more complex, i.e., two-dimension DCT basis vector, and (b) a low value when the block has less complex.

3.3.4 Sum of Absolute AC Coefficient Attack (SAC)

SAC generates a sketch image ϕ_S by the sum of absolute AC coefficients. Specifically,

$$\phi_S(i,j) = \text{round}\left(255 \times \frac{s(i,j)}{\max\{s(i,j)\}}\right). \tag{3.5}$$

where the sum of absolute of AC coefficients in the (i,j)-th 4×4 block is denoted as $s(i,j)$.

3.3.5 MB Bitstream Size Attack (MBS)

In H.264/AVC encoders, an entropy coding (either CAVLC or CABAC) is utilized to encode the DCT coefficients of an MB. It is found that a larger number of bits is spent for an MB with higher spatial activity, and vice versa. Therefore, the complexity of an MB infers as the bitstream size, which varies according to H.264/AVC compressed videos. An example is shown in Figure 3.1, which shows four MBs (labeled as A, B, C and D)

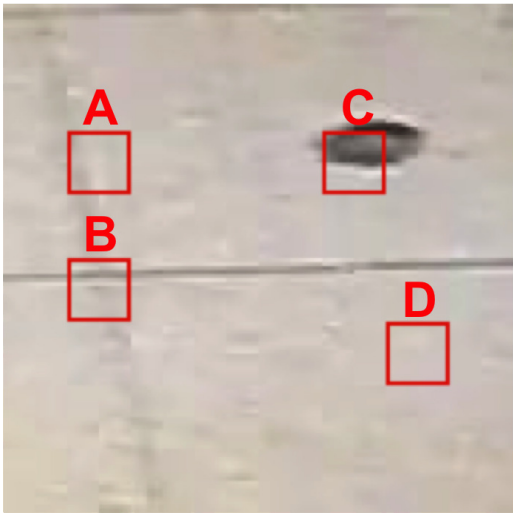


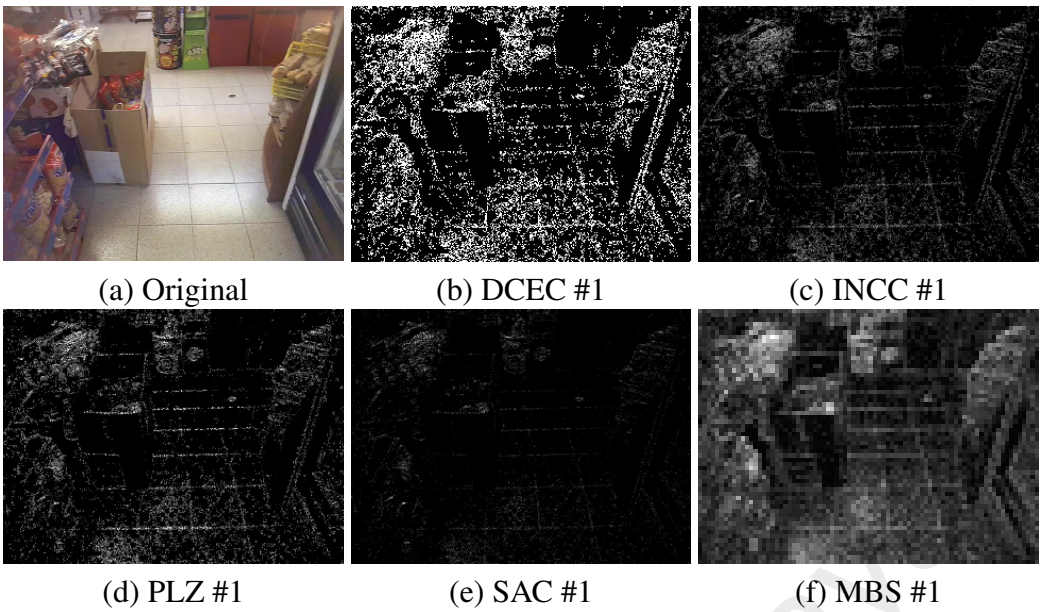
Figure 3.1: Four MBs (regions) with different spatial activities in a video frame compressed by using H.264/AVC

in a 128×128 pixels region of an H.264/AVC compressed video frame. The marked MB are different spatial activities. In particularly, 80 bits are allocated to region A consists of a weak diagonal edge. On the other hand, 208 bits are allocated to region B consisting of borders of the floor tiles, i.e., edges. Meanwhile, 146 bits are allocated to region C including two areas of different colors along with an edge separating them. Finally, only 24 bits are allocated to smooth region D with no edges. Based on these observations, an MB-based sketch image, viz., MBS image ϕ_B, can be generated as follows:

$$\phi_B(i_m, j_m) = \text{round}\left(255 \times \frac{b(i_m, j_m)}{\max\{b(i_m, j_m)\}}\right), \tag{3.6}$$

where $i_m \in [0, X/16]$, $j_m \in [0, Y/16]$, hence the resolution is $1/16$ of its original counterpart. The number of bits spent on encoding the (i_m, j_m)-th MB is denoted as $b(i_m, j_m)$. Note that the ratio of $b(i_m, j_m)$ over $\max\{b(i_m, j_m)\}$ is considered by MBS.

In this section, sketch images by the proposed sketch attacks are presented. Specifically, the proposed sketch attacks are applied to an H.264/AVC compressed video. Figure 3.2 shows the original frame and the sketch images obtained by applying the proposed sketch attacks (viz.: DCEC; INCC; PLZ; SAC, and; MBS of the first frame of



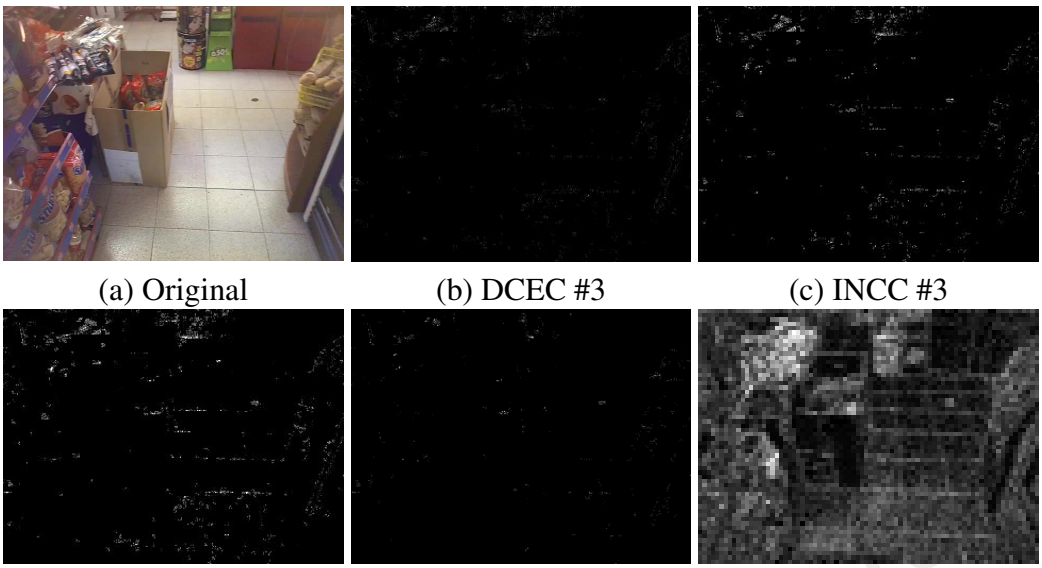
(a) Original (b) DCEC #1 (c) INCC #1

(d) PLZ #1 (e) SAC #1 (f) MBS #1

Figure 3.2: Original frame and sketch images of the original I-frame (in *Video 17* from ICDAR2013) generated by using: DCEC; INCC; PLZ; SAC, and; MBS

Video 17 from International Conference Document Analysis and Recognition 2013 (ICDAR2013)). It is found that outline images for the 1st frame (i.e., I-frame) are generated by the five proposed sketch attacks. On the other hand, the original frames and the sketch images for the third frame (i.e., P-frame) are shown in Figure 3.3. Result suggests that, when handling an INTER-frame such as P- or B-frame, Coefficient-based sketch attacks, (viz.: DCEC; INCC; PLZ, and; SAC) fail to generate a perceptually meaningful sketch image.

The reason is that only the information of the I-frame is encoded independently from the rest of the frames (similar to the scenario of still-image), while the information of the differences between the current and previous/future frames through motion estimation is encoded for P- or B-frames. It is concluded that the coefficient-based sketch attacks are less effective in sketching H.264/AVC compressed videos, when compared to JPEG (B.Pennebaker & L.Mitchell, 1992) compressed images. On the other hand, regardless of the frame type, the sketch images by MBS are viable to show outline information of the H.264/AVC compressed video.



(a) Original (b) DCEC #3 (c) INCC #3

(d) PLZ #3 (e) SAC #3 (f) MBS #3

Figure 3.3: Original frame and sketch images of the original P-frame (in *Video 17* from ICDAR2013) generated by using: DCEC; INCC; PLZ; SAC, and; MBS

3.4 Sketch Image Evaluation

Recall that Li et al.'s (W. Li & Yuan, 2007) evaluation method (see Section 2.4.3) is designed to map one output value for 8×8 pixel block in constructing the NZCA image. Therefore the evaluation is not applicable directly to H.264/AVC compressed video. To evaluate sketch images in terms of quantity, two steps: (a) binary edge image definition, and; (b) edge similarity evaluation, are considered by extending the evaluation method of NZCA to different block sizes.

3.4.1 Definition of Reference Binary Edge Image

In the literature, there are many edge detectors (see Section 2.4.1), including CAN (Canny, 1986), SOB (Gonzalez & Woods, 2006), and LOG (Gonzalez & Woods, 2006). To obtain the reference edge map Γ, the CAN is adopted as the edge detector (instead of SOB used by Li et al.), because hysteresis thresholding, which can detect continuous edge curve, is employed by the CAN and even some weak parts of an edge can be detected by the CAN. A sketch image ϕ depends on the information of a block transformation or an MB. In other words, a sketch image ϕ is of size $X/\delta \times Y/\delta$ for δ being the size of a block

transformation or an MB. Thus, ϕ is smaller than the CAN edge image Γ which is of size $X \times Y$. Therefore, to match the dimension of the sketch image, resizing the Canny binary edge map Γ is required. To generate the down-scaled CAN binary edge map Γ_D, a few operations (i.e., block partition, edge candidate calculation, and edge candidate classifying) are conducted. First, non-overlapping blocks each of $\delta \times \delta$ pixels are obtained by dividing the CAN binary edge map. Next, by using following the equation, the number of edge candidates for each block Γ_N is calculated. The equation is as follows:

$$\Gamma_N(i,j) = \sum_{u=1}^{\delta} \sum_{v=1}^{\delta} \Gamma_{u,v}(i,j). \tag{3.7}$$

Here, $\Gamma_{u,v}(i,j)$ denotes the (u,v)-th value in the (i,j)-th $\delta \times \delta$ block, where $1 \leq i \leq X/\delta$, $1 \leq j \leq Y/\delta$ and $1 \leq u,v \leq \delta$.

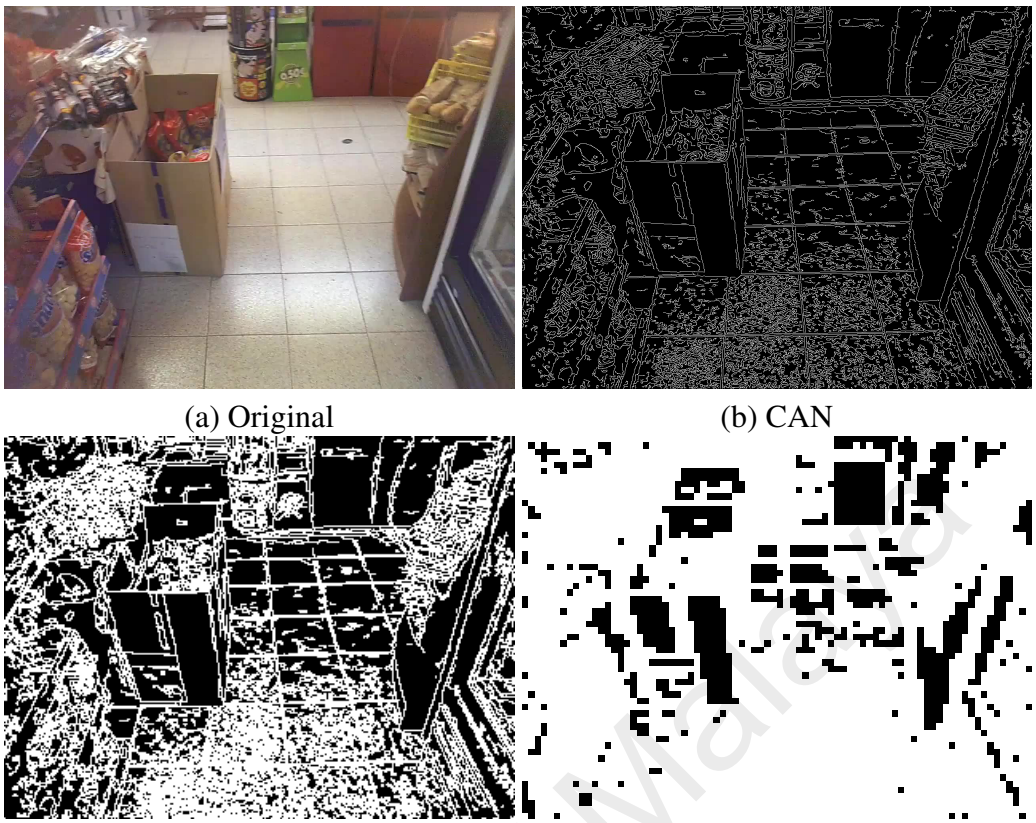
The number of edge candidates for each block is classified to generate the down-scaled CAN binary edge map Γ_D as follows:

$$\Gamma_D(i,j) = \begin{cases} 1 & \text{if } \Gamma_N(i,j) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

The original image, the CAN edge map and the down-scaled CAN edge maps are shown in Figure 3.4.

3.4.2 Edge Similarity Score (ESS)

The obtained down-scaled CAN edge map Γ_D is utilized in the edge similarity metric (W. Li & Yuan, 2007) (see Section 2.4.3) instead of SOB edge map Γ_0. In addition, the negated down-scaled edge images of Figure 3.4(c) and (d) are shown in Figure 3.5(a) and (b), respectively. It is observed that both outlines of the down-scaled images and their negated images are almost similar. In addition, for edge similarity evaluation, only the edges are of interest regardless of whether the sketch image is positively or negatively



(a) Original (b) CAN

(c) Down-scaled ($\delta = 4$) (d) Down-scaled ($\delta = 16$)

Figure 3.4: Illustration of CAN edge map and down-scaled CAN edge maps

(a) Negated down-scaled ($\delta = 4$) (b) Negated down-scaled ($\delta = 16$)

Figure 3.5: Illustration of negated down-scaled CAN edge maps

correlated. Therefore, the absolute value of Equation (3.9) is considered in this study. The defined ESS, ζ is denoted as follows:

$$\zeta = |\zeta_0| = \left| \frac{\eta_1^{'}}{2\Delta} + \frac{\eta_2^{'}}{2\Theta} - \frac{\eta_3^{'}}{2\Delta} - \frac{\eta_4^{'}}{2\Theta} \right|. \tag{3.9}$$

where Δ and Θ denote the number of 0's and 1's in the down-scaled CAN edge map Γ_D, respectively. In addition, $\eta_1^{'}$, $\eta_2^{'}$, $\eta_3^{'}$, and $\eta_4^{'}$ are defined as follows:

0.3406 0.5303 0.4016 0.2594 0.3151

(a) *Video 17* in ICDAR2013

0.4352 0.4037 0.5883 0.6357 0.6011

(b) *Video 20* in ICDAR2013

Figure 3.6: Sketch images using MBS for frame #1 to #5 and the corresponding ESS scores

$$\eta_1^{'} = |\{(i,j) : \Gamma_D(i,j) = 0 \ \& \ \phi(i,j) = 0\}|, \tag{3.10}$$

$$\eta_2^{'} = |\{(i,j) : \Gamma_D(i,j) = 1 \ \& \ \phi(i,j) = 1\}|, \tag{3.11}$$

$$\eta_3^{'} = |\{(i,j) : \Gamma_D(i,j) = 0 \ \& \ \phi(i,j) = 1\}|, \tag{3.12}$$

$$\eta_4^{'} = |\{(i,j) : \Gamma_D(i,j) = 1 \ \& \ \phi(i,j) = 0\}|. \tag{3.13}$$

Note that $\zeta \in [0,1]$, where any value nearer to unity indicates a better match between the sketch image ϕ and the down-scaled CAN edge map Γ_D, where CAN edge map is assumed as the ideal outline. In other words, a larger value implies better sketch quality, and vice versa. To demonstrate the sketch images and their obtained ESS, the MBS sketch images and their corresponding ESS are shown in Figure 3.6. It is observed that MBS is viable to sketch the outline of both steady (upper row, i.e., *Video 17* in ICDAR2013 - browsing products in a super market) and non-steady (lower row, i.e., *Video 20* in ICDAR2013 - captured from driver's seat) videos. The most ESS values of the non-steady video are higher than the corresponding ESS values of the steady video. The reasons are that non-steady video has more scene changes and the MBs need to more bits to en-

code. Therefore, MBS can effectively exploit this property to generate a better outline. In addition, it is observed that blurred sketch images yield lower ESS scores, and vice versa.

3.5 Experiments and Discussions

In this section, six video sequences, viz.: two video sequences from the ICDAR2013 video dataset (Karatzas et al., 2013), two video sequences from Ultra High Definition High Efficiency Video Coding (HEVC) dash dataset (Feuvre et al., 2014), and two video sequences from Xiph.org video test media (HD content and above) (*Xiph.org video test media*, n.d.), are considered to verify the viability of the proposed sketch attacks. Note that the resolution of the ICDAR2013 videos is 1280×960 pixels, while the resolution of the HEVC dash and Xiph.org video test media videos range from 1920×1080 to 3840×2160 pixels. To conduct experiments, using the BP with level 5.1, the first 30 frames of each test video are encoded into H.264/AVC format and then partially decoded for sketching the outline. To avoid loss of generality, the initial QP in H.264/AVC is set at 30. To implement the proposed sketch attacks, the H.264/AVC reference software (Tourapis Michael et al., 2009) is utilized for partial decoding purpose to obtain the components of a H.264/AVC compressed videos, then the sketch images are generated by implementing codes on Matlab. Note that, when an MB of type Pulse-Code Modulation (PCM) is encountered, the value of the 2nd largest MB bitstream size within the same frame will be output. Last, to obtain the final binary outline image, Otsu thresholding (Otsu, 1979) is applied to the obtained sketch image.

3.5.1 Non-Encrypted Video

Generally, the two types of frames (i.e., INTRA- and INTER-frame) are utilized to form an H.264/AVC compressed video, therefore, the two types of frame are considered separately. In this sub-section, to show the performance of the proposed sketch attacks, the non-encrypted video (i.e., the best case scenario) is considered. This case scenario is

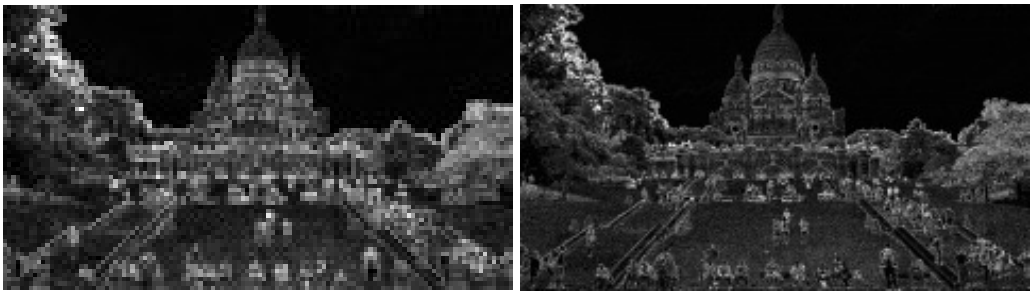
Table 3.1: ESS of sketch images for H.264/AVC INTRA-frame with initial QP $= 20, 30,$ and 40

Dataset	Sequence	QP	DCEC	INCC	PLZ	SAC	MBS
ICDAR 2013	Video 17 (1280×960)	20	0.1884	0.4062	0.3595	0.2140	**0.4595**
		30	0.2194	0.1531	0.2664	0.1684	**0.3406**
		40	0.0436	0.0345	0.0345	0.0345	**0.3162**
	Video 20 (1280×960)	20	0.1851	0.5795	**0.6194**	0.2707	0.5210
		30	0.1665	0.3527	0.3686	0.2193	**0.4352**
		40	0.1186	0.1233	0.1233	0.1233	**0.4524**
HEVC dash	v5 (1920 × 1080)	20	0.0920	0.4653	0.4881	0.2092	**0.7494**
		30	0.1340	0.2426	0.2946	0.1491	**0.5001**
		40	0.0947	0.0394	0.0750	0.0409	**0.4549**
	v9 (3840 × 2160)	20	0.0901	0.4912	0.5660	0.1559	**0.8062**
		30	0.1032	0.2346	0.2664	0.1033	**0.4209**
		40	0.0650	0.0608	0.0464	0.0608	**0.3780**
Xiph	Old town cross (3840×2160)	20	0.2503	0.3228	0.2748	**0.3798**	0.3299
		30	0.2910	0.3576	0.2910	0.1490	**0.6857**
		40	0.0421	0.0204	0.0531	0.0204	**0.3989**
	Rush hour (1920×1080)	20	0.1413	0.1772	0.1090	0.1076	**0.3853**
		30	0.1235	0.0606	0.0606	0.0606	**0.4914**
		40	0.0162	0.0035	0.0169	0.0035	**0.1876**

crucial to show the performance of the proposed sketch attacks, because if the proposed sketch attacks cannot even sketch the outline from the plaintext video, it will not be able to sketch from the encrypted video. Since sketch attack exploits the features (i.e., the complexity of each MB) in the compressed domain, if the complexity features are similar in both a non-encrypted video and its encrypted video, the sketch images obtained from the non-encrypted and encrypted videos are similar.

3.5.1.1 INTRA-Frame

When the sketch images by five sketch attacks on the INTRA-frames of six videos (i.e., two videos from every dataset, three datasets in total) are considered, the ESS values recorded in Table 3.1 are obtained, where the best ESS result among five sketch attacks is bolded for quick look-up purpose. It is observed that, when QP increases regardless of the video in question, except for *Old town cross* and *Rush hour*, the ESS values for all sketch attacks decrease. Besides, when QP ≤ 30, MBS yields the best results. In



(a) Sketch image generated from original video frame with resolution 1920×1080 pixels

(b) Sketch image generated from original video frame with resolution 3840×2160 pixels

Figure 3.7: Sketch images by MBS for the same video (UHD HEVC Dash Dataset) with two different resolutions

addition, when compared to all considered conventional methods for QP = 40 (i.e., the low bitrate case), the sketch images of MBS are particularly high ESS values. Hence, MBS appears to be robust against heavy compression (i.e., using large QP value). The reason is that most DCT coefficients both AC and DC) are quantized to zeros to reduce redundant information in the compressed videos, hence, in terms of extracting information from the compressed videos, the coefficient-based sketch attacks (i.e., DCEC, INCC, PLZ and SAC) are ineffective. Therefore, MBS is viable to generate sketch images at a quality comparable to that of the other sketch attacks for smaller QPs when handling INTRA-frame. In the rest case of QP $\geq$ 30, the ESS values of MBS further outperform.

In addition, it is observed that, when the resolution increases, the quality of the sketch images improves. To verify the trend, the sketch images of the same video frame at different resolutions are shown in Figure 3.7(a) and (b), where two difference resolutions, i.e., 1920×1080 and 3840×2160 pixels are considered. It is an expected result that generating a sketch image of high-resolution gives a high-quality sketch image, because encoding high-resolution frame requires more MBs, which include more semantic information of the frame. Therefore, when scaled to the same size for display, sketch images of higher resolution will show greater detail, and vice versa.

Table 3.2: Average ESS of sketch images for H.264/AVC INTER-frame with initial QP $= 20, 30,$ and 40

Dataset	Sequenc	QP	DCEC	INCC	PLZ	SAC	MBS
ICDAR 2013	Video 17 (1280×960)	20	0.0420	0.1987	0.2044	0.0886	**0.4220**
		30	0.0432	0.1779	0.1871	0.0808	**0.4115**
		40	0.0376	0.1372	0.1442	0.0684	**0.3841**
	Video 20 (1280×960)	20	0.0235	0.0330	0.0177	0.0444	**0.5490**
		30	0.0219	0.0317	0.0162	0.0421	**0.5469**
		40	0.0206	0.0302	0.0178	0.0396	**0.5298**
HEVC dash	v5 (1920×1080)	20	0.0001	0.0005	0.0004	0.0002	**0.2636**
		30	0.0003	0.0006	0.0006	0.0003	**0.3061**
		40	0.0012	0.0018	0.0016	0.0009	**0.3357**
	v9 (3840×2160)	20	0.0002	0.0006	0.0004	0.0004	**0.5671**
		30	0.0003	0.0007	0.0006	0.0004	**0.5668**
		40	0.0019	0.0027	0.0027	0.0025	**0.5496**
Xiph	Old town cross (3840×2160)	20	0.0293	0.0288	0.0139	0.0306	**0.4407**
		30	0.0205	0.0210	0.0136	0.0234	**0.4059**
		40	0.0202	0.0267	0.0202	0.0261	**0.3940**
	Rush hour (1920×1080)	20	0.1156	0.0876	0.1223	0.0711	**0.3337**
		30	0.1052	0.0834	0.1163	0.0623	**0.3321**
		40	0.0975	0.0787	0.1051	0.0566	**0.3366**

3.5.1.2 INTER-Frame

Next, the average ESS values (see Table 3.2) for the case of INTER-frame are discussed. It is observed that, regardless of the video and QP value under consideration, MBS always yields, by far, the highest ESS values. Results suggest that MBS is superior and capable of extracting information from INTER-frames where much of the redundancy is removed. Therefore, when compared to other proposed sketch attacks, MBS sketch attack is consistent and more robust to frame-type and QP parameter.

3.5.2 Format-Compliant Selectively Encrypted Video

To proof the viability of sketch attacks, two recently published format-compliant selective encryption methods (Yongsheng et al., 2013; B. Zeng et al., 2014) are considered in this section. Since, in Section 3.5.1, the sketch images of DCEC gave the low performance, only four sketch attacks, namely INCC, PLZ, SAC, and MBS, are henceforth considered.

As the representative test video sequences, 30 frames from the videos in the IC-

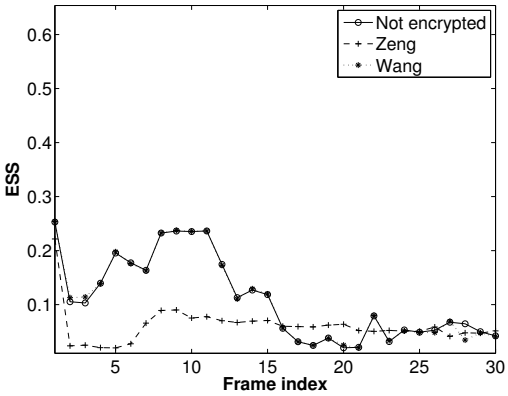


Figure 3.8: ESS of INCC for various encryption methods

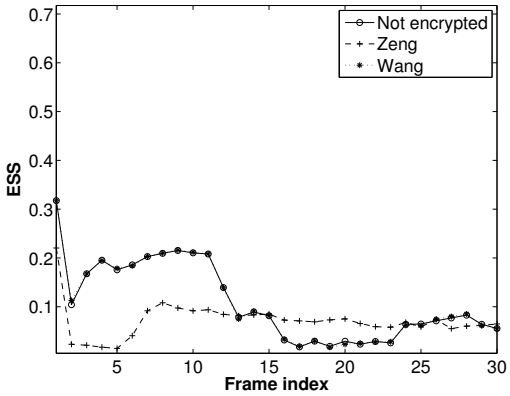


Figure 3.9: ESS of PLZ for various encryption methods

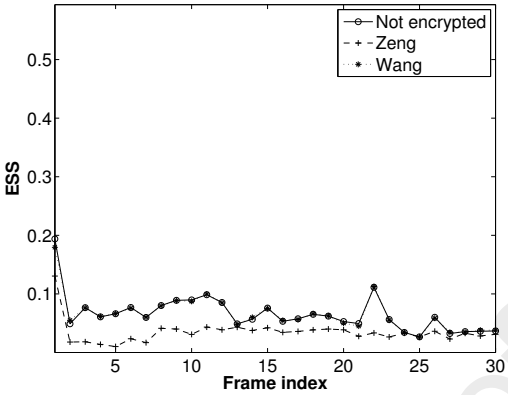


Figure 3.10: ESS of SAC for various encryption methods

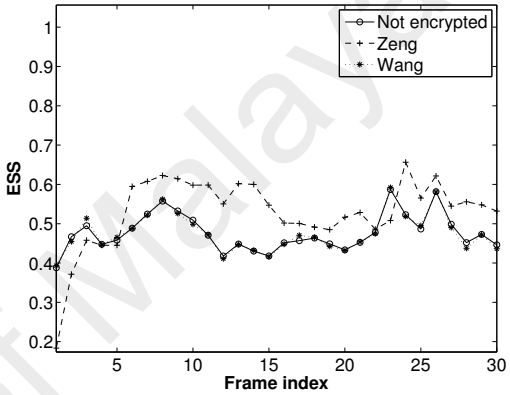


Figure 3.11: ESS of MBS for various encryption methods

DAR2013 dataset are utilized and then encrypted by using (Yongsheng et al., 2013) and (B. Zeng et al., 2014). The graphs of ESS for INCC, PLZ, SAC and MBS are shown in Figure 3.8, 3.9, 3.10, and 3.11, respectively.

Note that the *Not encrypted* curve refers to the obtained ESS score when comparing the ground truth (i.e., Canny binary edge map Γ_D) to the sketch images by the respective sketch attacks on the plaintext (i.e., not-encrypted) frames. It is observed that the ESS value for Wang et al.'s encryption method (Yongsheng et al., 2013) is almost identical to that of the original plaintext video. This trend suggests that the complexity of each MB, which could be exploited to generate sketch images of the original video, can not be masked by the encryption (Yongsheng et al., 2013). On the other hand, Zeng et al.'s method (B. Zeng et al., 2014) yields different trends for all considered sketch images. The trend includes higher and lower ESS values comparing with the counterparts of orig-

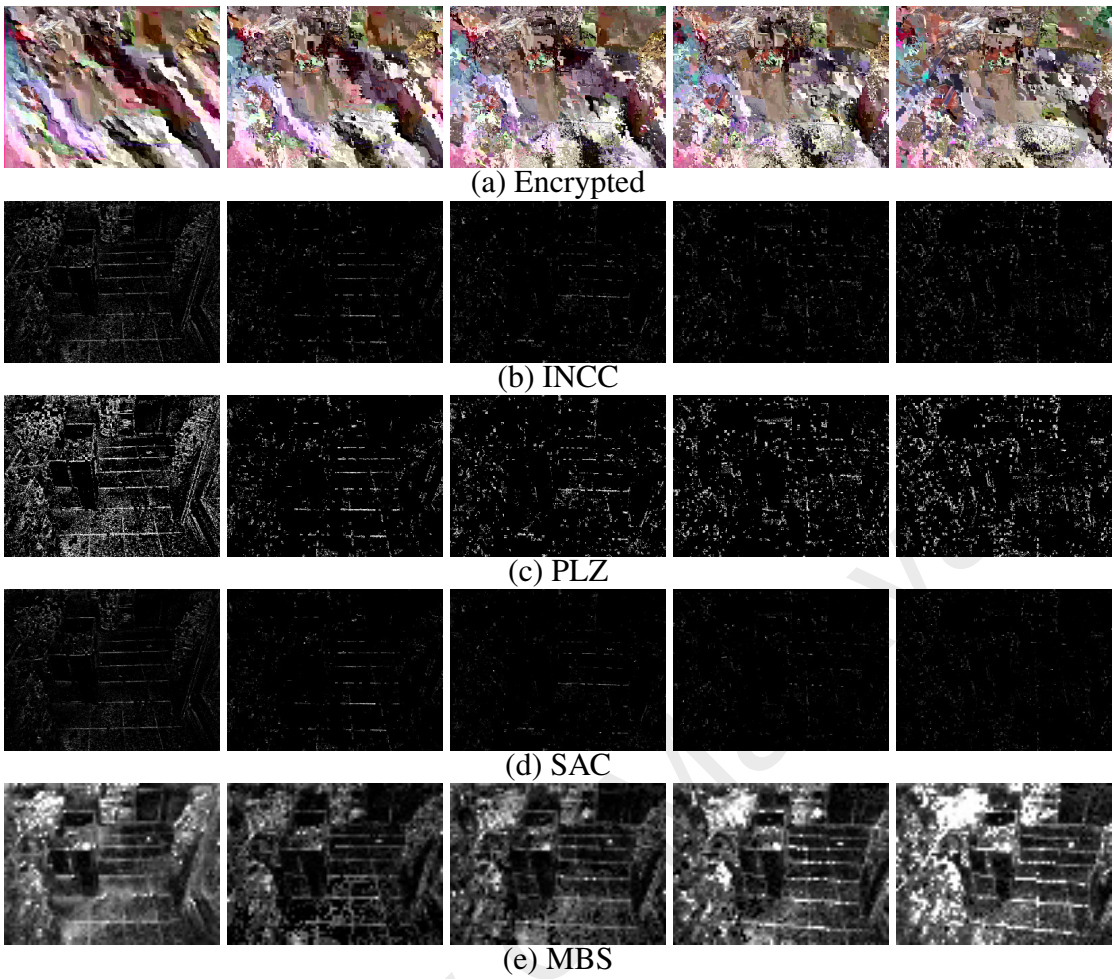


Figure 3.12: *Video 17* in ICDAR2013 (first 5 frames) encrypted by using Reference (B. Zeng et al., 2014) and the corresponding sketch images (2nd to 5th rows)

inal plaintext videos. A possible reason for this trend is that Zeng et al.'s encryption method (B. Zeng et al., 2014) changes the magnitude of coefficients, the utilized frequencies of coefficients and the number of coefficients, hence each MB bitstream size also changes. More noteworthy is that varied ESS values for all considered sketch attacks are observed, but nonetheless the highest ESSs are obtained by the MBS.

Next, the encrypted images by Zeng et al.'s method (B. Zeng et al., 2014) and the corresponding sketch images by all considered sketch attacks are demonstrated in Figure 3.12 and 3.13. Specifically, the first 5 frames of a representative steady video and a representative non-steady video (i.e., *Video 17* and *Video 20* from ICDAR2013, respectively) are considered. It is clearly observed that all considered sketch attacks are viable to generate sketch images of the first (i.e., INTRA-) frame, however, the three

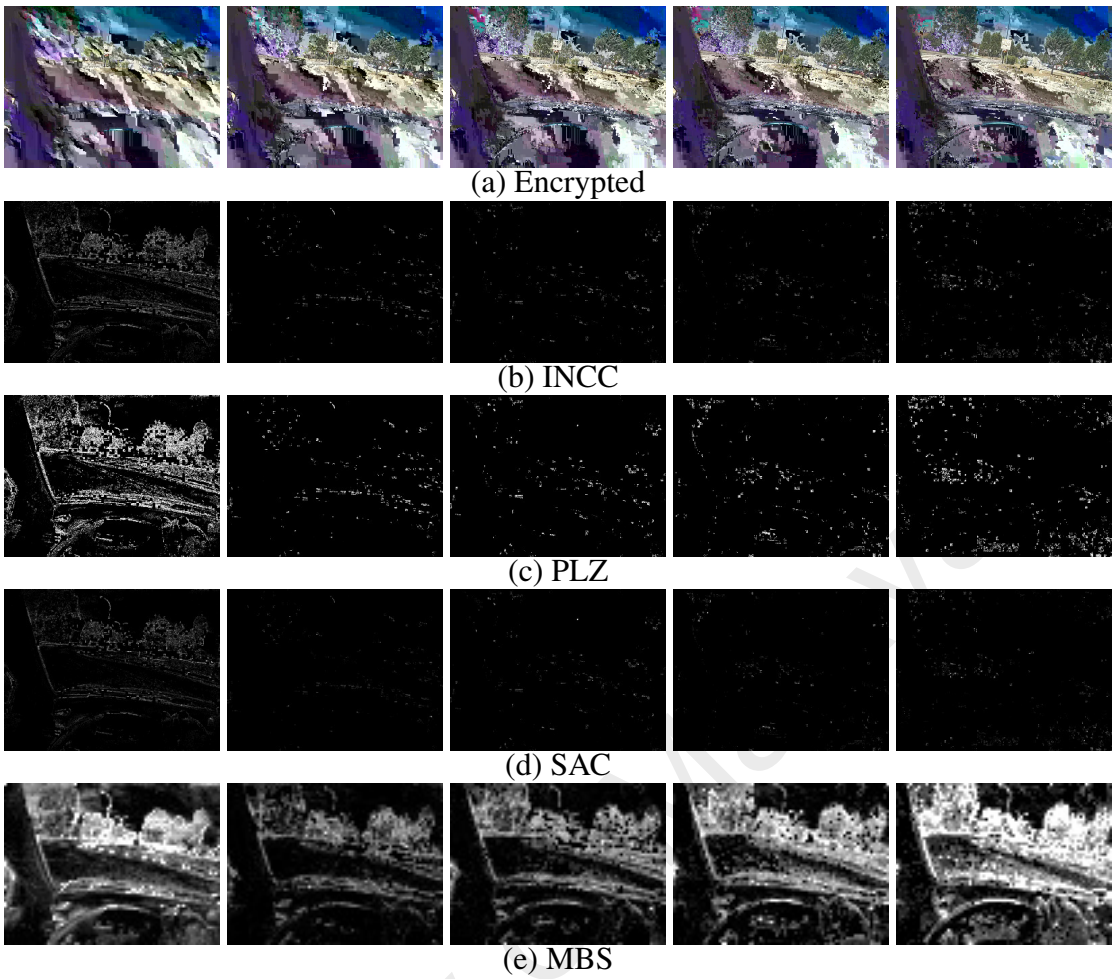


Figure 3.13: *Video 20* in ICDAR2013 (first 5 frames) encrypted by using Reference (B. Zeng et al., 2014) and the corresponding sketch images (2nd to 5th rows)

coefficient-based sketch attacks (i.e., INCC, PLZ, and SAC) failed to generate sketch images from INTER-frames while the MBS attack succeeded to generate the sketch images from INTER-frames. Besides, it is observed that the ESS curves for INTER-frames in Figure 3.11 are almost stable.

In addition, to demonstrate the ESS curves of all considered sketch attacks for various QPs, Figure 3.14 and Figure 3.15 respectively show the ESS graph of INTRA- and INTER-frames encrypted by Zeng et al.'s method (B. Zeng et al., 2014), where these frames are of *Video 17* from ICDAR2013 (Karatzas et al., 2013). In this case of INTRA-frame, the ESS values of the considered sketch images vary. Specifically, the ESS differences between MBS and other considered sketch attacks are small when QP is between 25 and 35, MBS always gives higher ESS than those of other considered sketch attacks

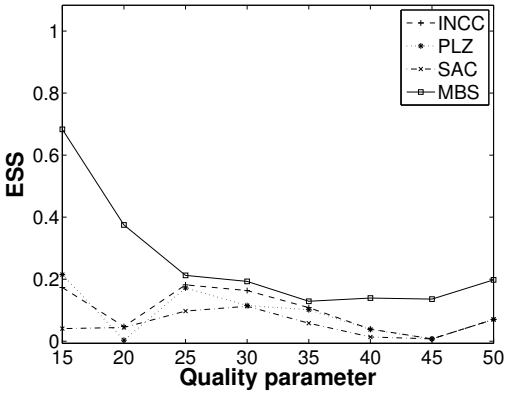


Figure 3.14: Graph of ESS vs. QP for various sketch attacks in INTRA-frame

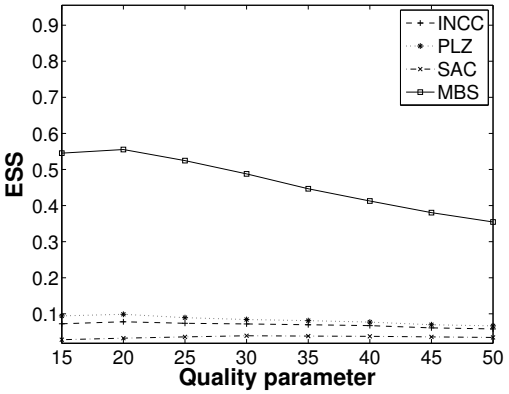


Figure 3.15: Graph of ESS vs. QP for various sketch attacks in INTER-frame

Table 3.3: Specific features exploited by each sketch attack

	Level_{DC}	Level_{AC}	Coef_{no}	Coef_{po}	MB bits
DCEC	✓	×	×	×	×
INCC	×	×	✓	×	×
PLZ	×	×	×	✓	×
SAC	×	✓	×	×	×
MBS	×	×	×	×	✓

for all QP's considered. On the other hand, since significant information reduction during video encoding for INTER-frames, ESS of INCC, PLZ, and SAC drastically dropped while MBS still yields stable performance with high ESS values. A similar trend is observed when considering other test videos. Therefore, it is concluded that, regardless of frame-type and QP, MBS is a viable sketch attack method.

3.5.3 Analysis

In this subsection, the considered sketch attacks are analyzed in terms of the exploited feature, viability to the encryption modules and the histogram of sketch image, respectively. First, Table 3.3 lists the exploited features by each sketch attack. Specifically, the items of the top row denote following features, respectively: (a) Level_{DC} is residual DC values, which is the outputs of prediction process in H.264/AVC; (b) Level_{AC} is the quantized AC coefficient values; (c) Coef_{no} is the number of nonzero AC coefficients of each block; (d) Coef_{po} is the position of the last nonzero AC coefficient index with respect to the Zigzag order in each block, and; (e) MB bits is the number of bits allocated to each

Table 3.4: Viability of sketch attacks for various H.264/AVC format-compliant selective encryption modules for video

	S	L	ST	QPM	DF	MVD	SSO	INTER-MM	INTRA-MM
DCEC	✓	×	×	✓	✓	✓	✓	×	✓
INCC	✓	✓	×	✓	✓	✓	✓	×	✓
PLZ	✓	✓	×	✓	✓	✓	×	×	✓
SAC	✓	×	×	✓	✓	✓	✓	×	✓
MBS	✓	×	✓	✓	✓	✓	✓	✓	✓

MB. Since the first four sketch attacks in the first column, i.e., DCEC, INCC, PLZ and SAC, utilize the IntDCT coefficients, hence it is observed that the sketch attacks fail to generate outline images when the IntDCT coefficient values are changed, while MBS is able to generate an outline image of the frame.

Next, the viability of each sketch attack in generating the outline directly from the encrypted H.264/AVC video is discussed. From format-compliant selective encryption modules in Section 2.3.2, the exploited feature in Table 3.3 and the aforementioned experiment observations, the viability of each sketch attack can be derived as shown in Table 3.4. Note that the interpretation of the Table is the same manner as in Table. 2.2. Specifically, the box of row α and column β is marked with ✓ $(\times)$, it means the sketch attack method α can (cannot) generate the outline image of the video when module β is applied to the video encryption method. Table 3.4 suggests that MBS is viable to generate an outline image directly from the encrypted video, except when IntDCT coefficient level modification is involved. Here, INCC and PLZ are robust to IntDCT coefficient level modification. This fact also suggests that appropriate sketch attack(s) can be selected by using some information about the encryption in use to obtain better sketch image(s) instead using one attack. For example, statistical analysis can be performed on the distribution of coefficient values, motion vectors, scanning orders, etc., for determining encryption modules, and then the appropriate attack is launched. This appropriate attack is beyond the scope of this research and the issue will be pursued in this direction as future

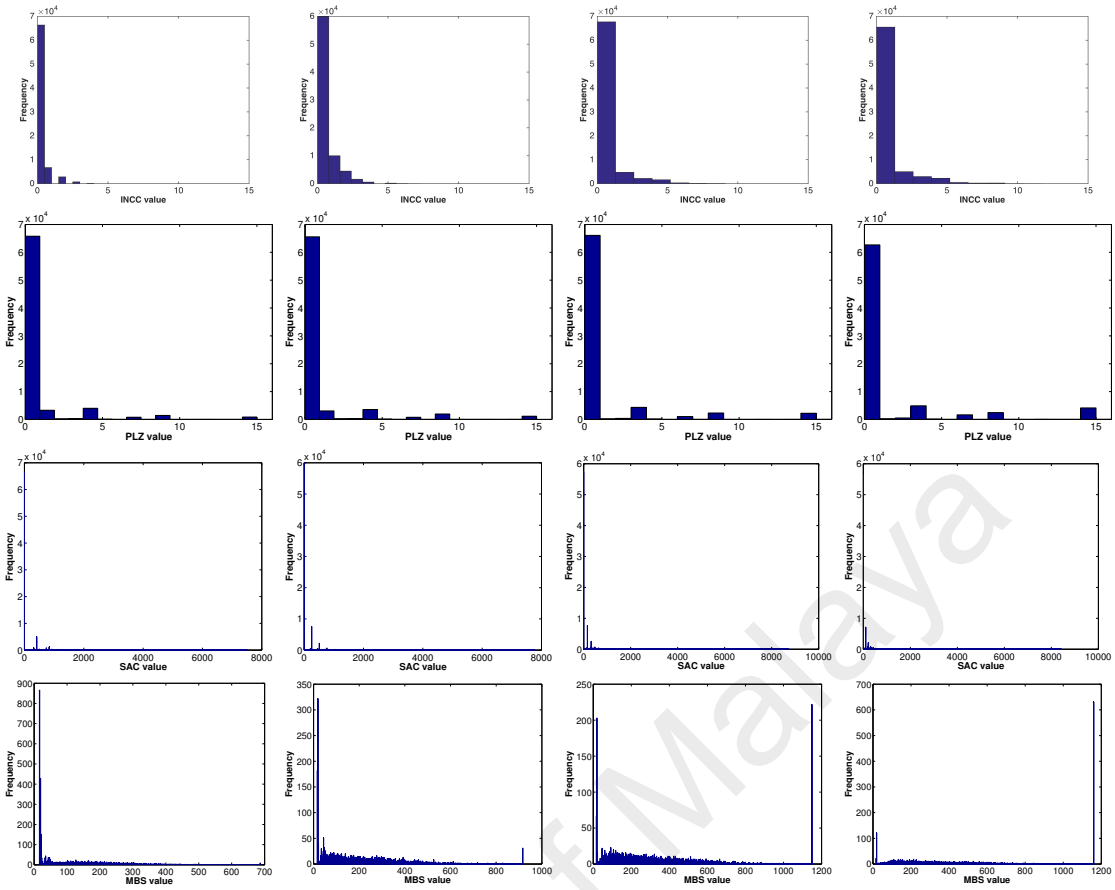


Figure 3.16: Histogram of each sketch images prior to applying Otsu's binarization for frame #2 to #5 shown in Figure 3.12

work.

Last, the histograms of the captured values are considered to clearly differentiate the information/statistics captured by each sketch attack. The obtained histograms of the sketch images by the considered sketch attacks are shown in Figure 3.16, where the plaintext video - *Video 17* from ICDAR2013 for frame #2 to frame #5 is considered as a test video. Here, the frame #1 (i.e., INTRA-frame) is not considered, because the sketch images by INCC, PLZ, SAC, and MBS give similar ESS values for INTRA-frame.

The theoretical range of bin values for INCC and PLZ is $[0, 15]$, but the observed ranges are smaller than the theoretical range, e.g., $[0, 7]$ in the case of INCC for frame #5 (see right-top in Figure 3.16), while the range for SAC bins is significantly wider. Generally, the range of bins has effect displaying an image/frame, hence, it is expected that the range for SAC displays a more contrasted image than those of INCC and PLZ. How-

ever, since the hight compression efficiency of H.264/AVC, the bins of SAC are mostly concentrated at the 0-th bin. Thus, low contrast images are obtained (see Figure 3.12).

On the other hand, only the distribution of MBS histogram is widely spread among all the histograms considered. The fact suggests that MBS can generate outlines of higher entropy (i.e., greater detail) when compared to the other sketch attacks. Therefore, a high contrast sketch image can be obtained by MBS.

In addition, Otsu's binarization method is applied to the sketch images (i.e., the INCC, PLZ, SAC and MBS) for obtaining binary images. Note that Otsu's binarization method utilizes bi-modal histogram properties to find the optimum threshold value. Hence, since the histograms of INCC, PLZ, and SAC are mostly concentrated at the 0-th bin, most pixel values will be treated as background, in other words, the intensity of pixels is zero.

3.5.4 Is MBS Viable?

As the aforementioned discussions in Section 3.5.3, it is empirically found that MBS is an effective attack in generating outline image from both INTRA- and INTER-frames encrypted by the conventional H.264/AVC format-compliant selective encryption modules. However, since MBS is designed to exploit each MB information (i.e., the bits spent to encode each non-overlapping 16×16 pixel block), straightforwardly, a shuffling MB operation with maintaining format-compliant can be implemented to withstand the proposed MBS sketch attack. It is because that MBS generates a pixel value from the corresponding MB. Thus when the MBs of a video frame is shuffled, the generated pixel values will also be shuffled. In other words, the shuffling operation prevents leaking the information of the plaintext video. Although the shuffling operation before or during video compression breaks the correlation among MBs, the operation induces the bitstream size overhead. Hence researchers refrained from deploying the shuffling operation for video

encryption. In addition, the shuffling operation requires additional operations, which are fully/partially decoding the compressed-encrypted video, to complete the decryption process. This requirement makes this encryption approach less feasible. On the other hand, a shuffling MBs operation during or after video compression can be considered. In H.264/AVC video compression, an error resilient tool of the Flexible MB Ordering (FMO), which allows a video frame to be arranged in different scan patterns of MBs, is included. Note that the FMO includes 6 built-in scan patterns as well as one option to include an entire explicitly user-defined pattern viz., assigned through the parameter MB Allocation MAP (MBAmap). Hence, the FMO information can be utilized to shuffle MBs. Specifically, by masking the scan pattern of MBAmap, the FMO-based shuffling can be deployed. However, FMO is only supported in the baseline and extended profiles of H.264/AVC (Wiegand et al., 2003), this approach is also less practical.

Another alternative encryption approach against MBS is diffusing the perceptual information of a region into other regions. For example, an encryption operation diffuses pixel values of an MB into other pixel values of MBs. Two diffusion scenarios using this approach can be considered, namely diffusion before or during video compression, and diffusion during or after video compression. In the first case, the diffusion approach destroys the spatial correlation among the pixels of the original frame, hence the compression efficiency will be dropped and the bitstream size overhead increases. In the second case, the diffusion can theoretically maintain format-compliant, however, since the nature of context adaptive entropy coding (i.e., CAVLC and CABAC) is complex, the diffusion is not able to be implemented straightforwardly. It is because the adaptive entropy coding uses multiple tables or statistical features of the input frame(s), hence the original statistical information is required to maintain bitstream size. In addition, the decode-encrypt-encode approach is always able to encrypt a compressed video, but the cost in terms of time and space complexities are too high, especially for smart devices

with limited battery power. It should be noted that the approach significantly causes the bitstream size overhead.

It can be concluded that, since the format compliance and bitstream overhead suppression are required for H.264/AVC video encryption, there are no straightforward encryption methods, which can avoid local complexity information leakage (i.e., the complexity of each MB) as well as the aforementioned requirements. Therefore, the proposed MBS sketch attack is viable to extract outline information of the original frame directly from the encrypted video in H.264/AVC format, which supports FMO in only a few profiles, i.e., baseline and extension. However, if security is the uppermost important issue for a particular video, then the user can consider the shuffling operation in FMO or the decode-encrypt-encode approach to withstand the proposed MBS at the expense of bitstream size overhead or high-cost computation and space. All in all, there is a trade-off relationship among security, format compliance, and bitstream size overhead.

3.6 Summary

In this chapter, five novel sketch attacks for format-compliant selectively encrypted video, namely encrypted H.264/AVC video, were put forward to generate sketch images, which can reveal an outline of each frame of the original video. Specifically, sketch images are generated using the DC, AC coefficients or the MB bitstream size of the encrypted video. In addition, to evaluate the performance of the sketch images by the proposed sketch attacks, an edge similarity method was modified and then the similarity scores were calculated by using the Canny edge map as an ideal outline image. It is observed from experimental results that the proposed MBS sketch attack can extract and reveal perceptual information directly from the video encrypted by format-compliant selective encryption. It should be emphasized that, although the five proposed sketch attacks can generate sketch images with a visible outline from the encrypted INTRA-frame, only the

MBS can extract outline from the encrypted INTER-frame, which outnumbers INTRA-frame, by far, in compressed video. Furthermore, it is validated that the MBS is more robust against various compression parameter (i.e., QP) than the performance of the rest proposed sketch attacks. Moreover, some future research directions are straightforward, such as launching appropriate sketch attack(s) by determining the encryption module in use, extending the sketch attacks to different video standards, i.e., HEVC, Audio Video Standard (AVS) and Google VP9.

CHAPTER 4: OUTLINE CLEARNESS ASSESSMENT

4.1 Overview

In this chapter, a no-reference outline clearness assessment metric is proposed as an outline quality assessment method. First, the incapability of the conventional no-reference image quality metrics in assessing outline images is shown. To address this problem, a novel no-reference outline clearness assessment metric is proposed that considers two properties, viz., image entropy and spatial correlation. To validate the feasibility of the proposed no-reference outline clearness assessment metric, experiments are conducted by using two image datasets, viz., the University of Southern California - Signal and Image Processing Institute (USC-SIPI) standard test images and the Berkeley Segmentation Dataset and Benchmark (BSDS500) images. Results suggest that the proposed outline clearness metric can give appropriate scores to outline images, and the scores are found to vary according to the observed clearness and distortion introduced. In addition, the proposed outline clearness metric is robust to compression and sensitive to Gaussian white noise.

4.2 Introduction

As overviewed in Section 2.4, the ideal outline cannot be obtained in practical situations and it is also futile to have a reference based outline clearness assessment metric. In addition, general assessment approach of IQA evaluates how similar the target image is with respect to the reference (original) image. On the contrary, the assessment approach of this study tries to evaluate the outline information extracted from the non-compressed or compressed data.

However, the features considered in conventional no-reference IQA metrics are features of a natural image, and the conventional edge similarity assessment (W. Li & Yuan,

2007) bases on binary images. Hence, the conventional assessment metrics may not be directly applicable to evaluate outline images. These shortcomings of the conventional assessment metrics motivated this study to design a no-reference clearness assessment metric for outline images.

Since not all detected outline images are useful to describe the outline information, the proposed assessment metric OCA makes an assumption, i.e., the ideal outline image has only plain regions and perceptible gradients, which consists of contrasted lines/textures. In other words, outline information depends on two factors: (a) image entropy (Gonzalez & Woods, 2006), which is a global feature indicating the amount of visual information based on the distribution of gray levels, and; (b) spatial correlation coefficient (Gonzalez & Woods, 2006), which is a local feature that captures the aggregated relation between each pixel and its neighboring pixels. The proposed assessment metric is designed to consider these two factors.

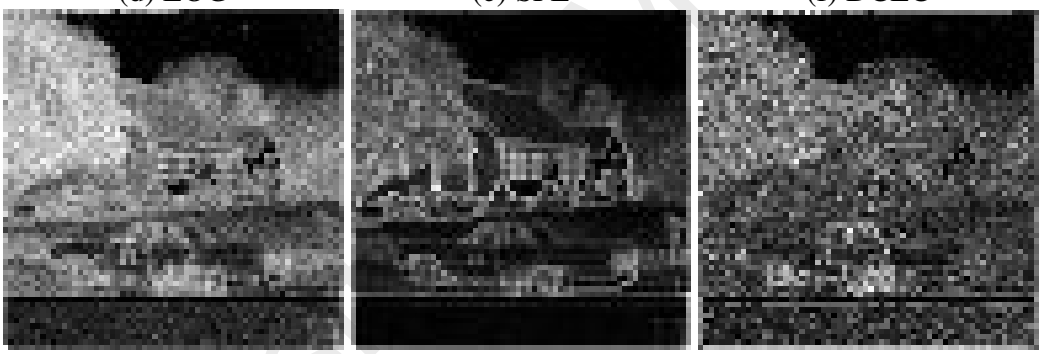
4.3 Limitations of Conventional No-reference IQA

Based on the literature survey carried out in this study, there are two types of outline detector, i.e., edge detector and sketch attack (see Section 2.4.1). In this Chapter, representative differentiation edge detectors, i.e., CAN (Canny, 1986), SOB (Gonzalez & Woods, 2006) and LOG (Gonzalez & Woods, 2006), and a representative learning based edge detector, SFE (Dollar & Zitnick, 2013) are considered. In addition, representative sketch attacks (i.e., DCEC, INCC, SAC, and MBS) are also considered. The original image and the corresponding outline images are shown in Figure 4.1.

Since the edge similarity assessment (W. Li & Yuan, 2007) bases on the SOB edge map of the original and IQA is designed for natural images, practical no-reference OCA metric is nonexistence. To evaluate the quality of outline images, a natural way is to apply a no-reference IQA metric, such as (Kamble & Bhurchandi, 2015; Lin & Jay

(a) Original (b) CAN (c) SOB

(d) LOG (e) SFE (f) DCEC



(g) INCC (h) SAC (i) MBS

Figure 4.1: Original image, and outline images generated by edge detectors (CAN, SOB, LOG, and SFE) and sketch attacks (DCEC, INCC, SAC, and MBS)

Kuo, 2011), directly. To investigate into the non-viability of this approach, the latest IQA metrics, viz., Spatial-Spectral Entropy-Based Quality (SSEQ) (L. Liu et al., 2014), Autoregressive-Based Image Sharpness Metric (ARISM) (Gu, Zhai, Lin, et al., 2015), Non-Reference Free Energy-Based Robust Metric (NFERM) (Gu, Zhai, Yang, & Zhang, 2015), and Multi-domain Structural and Global Frequency Features + Piecewise Regression (MSGF-PR) (Q. Wu et al., 2015), which predict the Difference Mean Opinion Score (DMOS) between the reference and the processed image, are considered.

Note that the scores range from $[0, 100]$, where DMOS $= 0$ implies the best quality

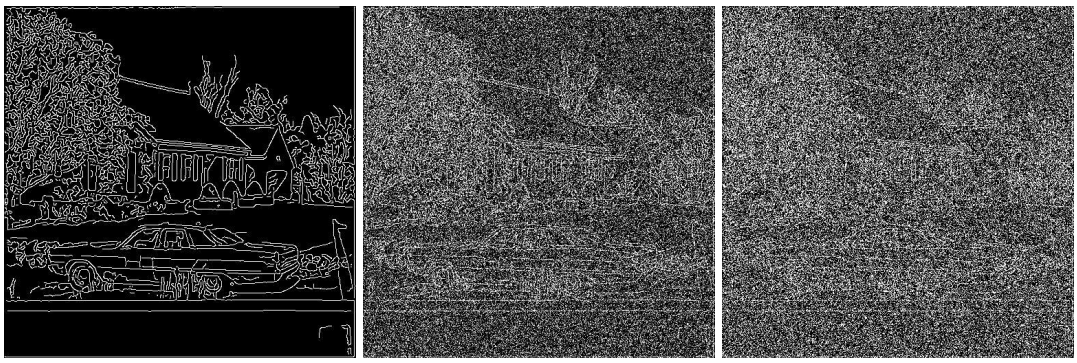
Table 4.1: Predicted DMOS using recent no-reference IQA metrics

Metric	CAN	SOB	LOG	SFE	DCEC	INCC	SAC	MBS
SSEQ	36.39	52.40	37.68	47.36	11.33	-2.13	8.78	37.29
ARISM	151.20	154.74	152.04	47.28	NaN	NaN	NaN	NaN
NFERM	162.04	133.16	167.79	55.17	6.63	36.35	9.21	77.48
MSGF-PR	-917.99	44.90	675.09	46.29	15.03	16.28	5.81	63.52

NaN indicates non-a-number

and DMOS = 100 implies the worst quality. The implementations of these IQA metrics are downloaded from the publicly accessible online repositories or obtained directly from the authors.

Table 4.1 records the predicted DMOS for the reference images after training with the Laboratory for Image and Video Engineering (LIVE) dataset (H.R. Sheikh & Bovik, n.d.). As a representative example, the outline image for *House* is shown in Figure 4.1, where the output of CAN, SOB and LOG (i.e., binary image) is mapped to grayscale image by using $0 \rightarrow 0$ and $1 \rightarrow 255$. It is observed that the scores in Table 4.1 deviate from my expectation. Specifically, by considering Figure 4.1, it is expected for CAN, SOB, LOG and SFE to yield better DMOS than DCEC, INCC, SAC, and MBS. However, DCEC or SAC appear to be the best approach to generate outline images among the approaches considered. Furthermore, SSEQ (L. Liu et al., 2014) and MSGF-PR (Q. Wu et al., 2015) yield negative values, which is out of their feasible ranges, and the result of ARISM (Gu, Zhai, Lin, et al., 2015) is undefined because the size of the sketch image is smaller than its minimum operational size. These observations imply that SSEQ, ARISM and MSGF-PR are incapable in evaluating the clearness of outline images. Moreover, NFERM (Gu, Zhai, Yang, & Zhang, 2015) yields larger DMOS values (i.e., worse result) for outline (edge) images generated by CAN, SOB, LOG, and SFE when compared to the outline (sketch) images generated by DCEC, INCC and SAC, which are counterintuitive. In addition, MSGF-PR (Q. Wu et al., 2015) may fail to assess noisy images with Additive Gaussian White Noise (AGWN) or salt and pepper noise. Figure 4.2 shows three outline



(a) $\sigma_A = 0$: 36.38 (b) $\sigma_A = 0.5$: 49.79 (c) $\sigma_A = 1.0$: 38.44

Figure 4.2: Canny outline images with various AGWN and the corresponding image clearness scores of MSGF-PR

images, each distorted with different AGWN. The standard deviation value σ_A in AGWN and the corresponding predicted DMOS using MSGF-PR are also recorded. It is observed that the MSGF-PR score is higher for the case of $\sigma_A = 1.0$ (see Figure 4.2(b)) when compare to the case of $\sigma_A = 0.5$ (see Figure 4.2(c)). This observation does not match the intuitive clearness assessment for the outline images shown in Figure 4.2.

Therefore, these results suggest that the conventional IQA metrics cannot be utilized directly as OCA because they do not consider the spatial correlations. These findings motivated this study to put forward a novel OCA metric.

4.4 Proposed OCA Metric

Literature survey suggests that is no OCA metric that could operate without referencing (i.e., no-reference) to the original reference image, and the conventional no-reference IQA metrics are found to be incapable in evaluating outline images (see Section 4.3). In that regards, measurable features of sketch image are considered. In this work, it is assumed that sketch image is not a color image, and it has outline and of high correlation in local regions. To capture these features, two measures are introduced for OCA, viz., Grayscale Image Information Entropy (GIIE), denoted by Λ_f, to capture the global characteristics, as well as Subsample Based SSIM (S3IM), denoted by Ξ_m, to capture the local character-

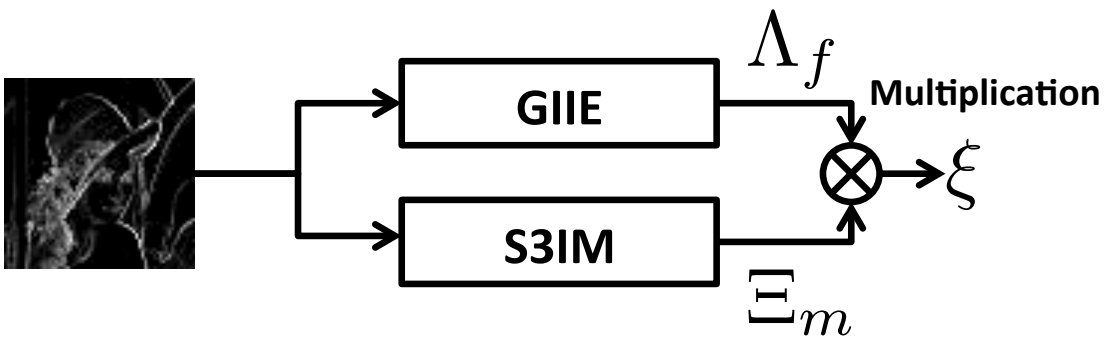


Figure 4.3: Process flow of the proposed OCA metric

istics. The product of these measures then forms the proposed OCA metric ξ. Figure 4.3 shows the processes involved in the proposed OCA metric.

4.4.1 Outline Criteria

There are two recent works (Hou et al., 2013; Lopez-Molina, De Baets, & Bustince, 2013) for edge evaluation based on full-reference edge images. The common conclusion drawn is that the ground truth labeled by human might be ambiguous and unreliable due to the influence of personal decision. In addition, although there are various sophisticated outline detectors (Liang & Yuen, 2013; Lopez-Molina, Baets, et al., 2013; McIlhagga, 2011; Yi et al., 2009), no-reference OCA does not exist, to date.

Since the ideal outline of an image may differ across applications, in this chapter, the global and local features of outline images are considered. Ideally, outline images should consist of some form of perceptual information, where its entropy must be greater than a certain value. Furthermore, the recently published IQA metrics (Guan et al., 2015; L. Li et al., 2015; F. Qian et al., 2015; Q. Wu et al., 2015) consider four types of distortion, viz.: (a) JPEG2000 compression; (b) JPEG compression; (c) AGWN, and; (d) Gaussian blur, for evaluation purpose. Among them, the distortion caused by (c) and (d) are more significant, because AGWN breaks the spatial correlation among neighboring pixels while Gaussian blur smoothens the image, viz., the overall contrast of the image decreases. Moreover, edge detectors (Lopez-Molina, Baets, et al., 2013) based on the

multi-scale approach achieve promising performance, because the multi-scale approach behaves similarly to the way humans interpret a scene. Based on these facts, it is assumed that significant outline information is retained at any scales. Therefore, the aforementioned two criteria, namely, image entropy for global feature analysis and subsample images correlation for local feature analysis are considered in the formulation of OCA metric.

4.4.2 Grayscale Image Information Entropy (GIIE)

Shannon (Shannon, 1948) quantifies information as entropy Λ by considering uncertainty in a random variable $\mathbb{Z}$ as follows:

$$\Lambda = -\sum_{z \in \mathbb{Z}} \rho(z) \log \rho(z), \tag{4.1}$$

where $z \in \mathbb{Z}$ is a random variable value, and $\rho(z)$ denotes the probability of z. In the case of 8-bit grayscale image (i.e., gray level in the range $[0, 255]$), the information entropy (GIIE) Λ_g is defined as follows:

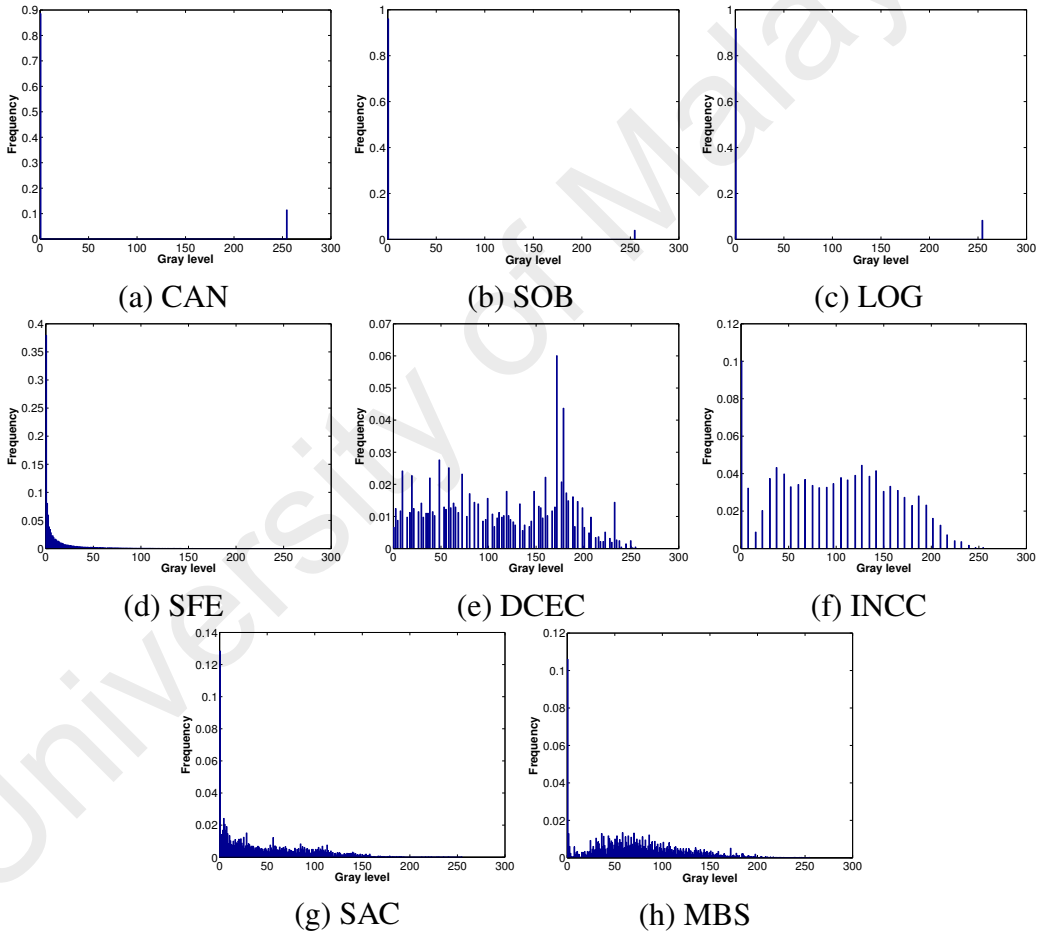
$$\Lambda_g = -\sum_{l=0}^{255} \rho(l) \log_2 \rho(l), \tag{4.2}$$

where l denotes the gray level and $\rho(l)$ denotes the probability of gray level l. The GIIE value can be considered as a dynamic range indicator for an outline image, where a higher value indicates a larger dynamic range, and vice versa. The GIIE value of an image close to the maximum value (e.g., 8 when considering 8-bit grayscale image) means nearly all gray levels are uniformly utilized to represent the image. By observation, a low entropy value will suffice for a visible outline.

Table 4.2 records the entropy for eight groups of outline images, where each group consists of nine standard test images (USC-SIPI, 2014), including Elaine, F-16, Fishing

Table 4.2: Entropy Λ_g for various outline images

Image	Pixel				DC	AC		Bitstream
	CAN	SOB	LOG	SFE	DCEC	INCC	SAC	MBS
Elaine	0.4351	0.1370	0.2478	3.7567	6.0554	4.1528	6.0846	6.8335
F-16	0.4070	0.2232	0.3196	3.8548	5.8432	4.6811	6.2806	7.1850
F.B.	0.5272	0.2427	0.4094	4.6556	5.9071	4.7391	6.7893	7.1994
House	0.5135	0.2408	0.4120	4.3956	6.0693	4.8163	6.7224	7.0287
Lenna	0.4242	0.2018	0.3026	4.1827	5.9679	4.4565	6.3459	7.0763
Mandrill	0.6875	0.2788	0.5299	4.0946	5.6853	5.0394	7.2762	7.3295
Peppers	0.3690	0.1721	0.2550	4.3476	6.1674	4.4022	6.1521	6.8074
S.L.	0.4479	0.2442	0.3674	4.6423	6.1051	4.8096	7.0608	7.0019
Splash	0.3613	0.1208	0.2143	3.3134	6.2088	4.0186	5.6770	7.1759
Average	0.4636	0.2068	0.3398	4.1381	6.0011	4.5684	6.4877	7.0708



(a) CAN (b) SOB (c) LOG

(d) SFE (e) DCEC (f) INCC

(g) SAC (h) MBS

Figure 4.4: Histogram of the original and outline images shown in Figure 4.1

Boat (F.B.), House, Lenna, Mandrill, Peppers, Sailboat on Lake (S.L.), and Splash. The first row indicates the metric used in generating the outline. The average entropy for CAN, SOB, and LOG yield lower values when compared to others, because the distribution of gray levels in these outline images are skewed (i.e., mostly zero to signify the non-edge

regions or background) and the edge responses are thresholded into two values. On the other hand, the rest of the edge detectors and the sketch attacks yield higher entropy. Figure 4.4 shows the histogram corresponding to each image shown in Figure 4.1. By observation, the histograms (of pixel values) of CAN, SOB, LOG and SFE outline images are skewed, but the outlines are still visible. Hence, it can conclude that a low entropy value will suffice for a visible outline.

To classify an outline image into either a visible or a non-visible outline image, a non-linear function Λ_f is introduced as follows:

$$\Lambda_f = \begin{cases} 1 & \text{if } \Lambda_g \geq \Lambda_t; \\ 0 & \text{otherwise,} \end{cases} \tag{4.3}$$

where Λ_t denotes a threshold value. In this Chapter, Λ_t is empirically set at 0.05, which yields the most suitable score matching to subjective assessment. Equation 4.3 classifies an outline image as a meaningful or meaningless outline image.

4.4.3 Subsample based SSIM (S3IM)

Recall that an assumption is made earlier where significant outline information is retained at any scales (see Section 4.4.1). In other words, the original and scaled outline images have similar outline structures. To quantify similarity in terms of structure, the Structural Similarity Index Measurement (SSIM) (Z. Wang et al., 2004) is modified in this Chapter. Since SSIM is a full-reference assessment, it cannot be applied directly to evaluate outline image without a reference image. Therefore, a modified SSIM, which transforms SSIM (i.e., full-reference metric) into a no-reference assessment metric by removing the reference image requirement, is proposed.

In the literature, multi-scale analysis (which analyzes images derived from the input image) is utilized in IQA metrics (J. Qian et al., 2014; Z. Wang et al., 2003) because

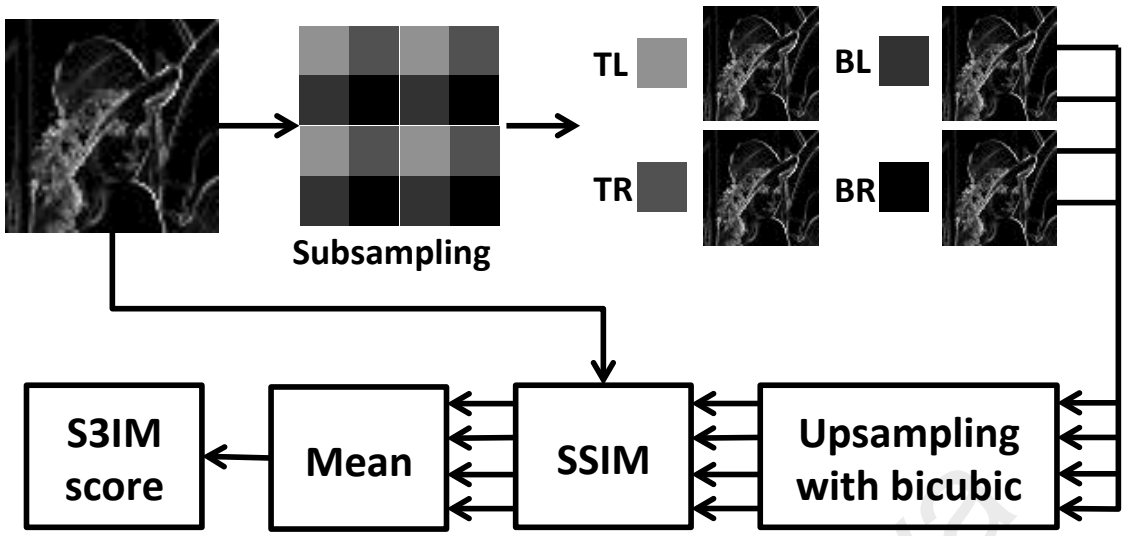


Figure 4.5: The concept of subsample based SSIM

multi-scale analysis resembles the mechanism of human visual system. Here, subsampled images are utilized for multi-scale analysis to retain the ability of SSIM (Z. Wang et al., 2004) in assessing structural information while removing the requirement of a reference image. Figure 4.5 shows the architecture overview of the proposed subsample-SSIM (S3IM) concept. First, an outline image of size $\mathbb{M} \times \mathbb{N}$ pixels is subsampled into 2×2 blocks. Within each 2×2 block, the pixels are labeled as Top-Left (TL), Top-Right (TR), Bottom-Left (BL) and, Bottom-Right (BR). Then, four subsampled images, each of size $\mathbb{M}/2 \times \mathbb{N}/2$ pixels, are formed by grouping the TL, TR, BL, and BR pixels. Next, each subsampled outline image is upsampled by using bicubic interpolation metric to attain the same size as the original outline image (i.e., $\mathbb{M} \times \mathbb{N}$). The upsampled images are denoted uTL, uTR, uBL and uBR, respectively. Then, four SSIM values (Z. Wang et al., 2004) are calculated for the following pairs: Original$\oplus$uTL, Original$\oplus$uTR, Original$\oplus$uBL, and Original$\oplus$uBR.

The SSIM values are correspondingly denoted by Ξ_{tl}, Ξ_{tr}, Ξ_{bl} and Ξ_{br}. Next, the average of these SSIM values is then calculated and output as the S3IM value Ξ_m. S3IM may be improved by subsampling at higher scales. However, the image considered (par-

Table 4.3: S3IM scores for grayscale images and their corresponding outline images

Image	Pixel				DC	AC		Bitstream
	CAN	SOB	LOG	SFE	DCEC	INCC	SAC	MBS
Elaine	0.7358	0.8168	0.6887	0.9759	0.6412	0.4638	0.5203	0.2972
F-16	0.7854	0.8148	0.7469	0.9737	0.6460	0.5435	0.6063	0.3461
F.B.	0.6783	0.7577	0.6329	0.9669	0.5827	0.5162	0.5223	0.3228
House	0.6945	0.7593	0.6677	0.9705	0.6417	0.5600	0.5872	0.3675
Lenna	0.7472	0.8117	0.7199	0.9688	0.6215	0.5367	0.5537	0.3638
Mandrill	0.6447	0.6928	0.5679	0.9788	0.5688	0.5400	0.5371	0.3642
Peppers	0.7576	0.8239	0.7414	0.9689	0.6467	0.5226	0.5542	0.3243
S.L.	0.7366	0.7411	0.6721	0.9691	0.5651	0.5352	0.5388	0.3419
Splash	0.7316	0.8873	0.7101	0.9770	0.7433	0.4478	0.5873	0.2550

ticularly sketch image) is relatively small, and hence much information will be removed when the image is further sub-sampled. This problem will be considered as future work.

Table 4.3 shows the S3IM values for the grayscale and the outline images. The S3IM values of SFE are higher when compared to other outline detectors. In addition, the S3IM values of SAC are the highest among AC coefficients based outline detectors. The results agree with the observation. With this construction, S3IM can measure the spatial correlation without the need of any reference images.

4.4.4 OCA Score

The proposed OCA metric ξ is defined by the product of two measures, viz., GIIE Λ_f and S3IM Ξ_m. The proposed OCA metric can be expressed as follows:

$$\xi = \Lambda_f \times \Xi_m, \tag{4.4}$$

where $\xi \in [0,1]$. A value nearer to unity indicates a better outline clearness, and vice versa. Hereinafter, the value calculated by the proposed OCA metric ξ is referred to as the *OCA score*.

Table 4.4: Average OCA scores for two image datasets

Dataset	Pixel				DC	AC		Bitstream
	CAN	SOB	LOG	SFE	DCEC	INCC	SAC	MBS
Standard	0.7235	0.7895	0.6831	0.9722	0.6286	0.5184	0.5564	0.3314
BSDS500	0.5342	0.7111	0.5151	0.9526	0.6325	0.5313	0.5322	0.3795

4.5 Experiment Results

To validate the proposed OCA metric, nine standard test images (USC-SIPI, 2014), including Elaine, F-16, F.B., House, Lenna, Mandrill, Peppers, S.L., Splash, as well as the BSDS500 (Arbelàez et al., 2011) for JPEG images are considered. Unless specified otherwise, the JPEG QF is set at 75.

4.5.1 OCA Score

Figure 4.6 shows the outline (edge and sketch) images generated by all eight outline detectors considered and records the corresponding OCA scores. The outline detectors include, in the order of left to right: (a) CAN, (b) SOB, (c) LOG, (d) SFE, (e) DCEC, (f) INCC, (g) SAC, and (h) MBS. The outline image obtained with SFE always yields the highest score among the outline detectors considered. In addition, the outline image by SOB achieves the second highest score. The outline image by MBS always attains the lowest scores because the outline images appear to be noisy. In the case of Mandrill (i.e., last row) the outlines by CAN, SOB and LOG yield relatively low scores when compared to other test images. It is because Mandrill is highly textured (e.g., consisting of hair and beard), and hence it is technically challenging to distinguish between noise and texture. Therefore, the proposed OCA yields lower scores for the outline images of Mandril generated by CAN, SOB and LOG.

Table 4.4 records the average OCA scores for the two image datasets considered, viz., standard test images and BSDS500. Results suggest that, for the BSDS500 dataset, the OCA scores of CAN are lower than those of DCEC. It is because BSDS500 includes

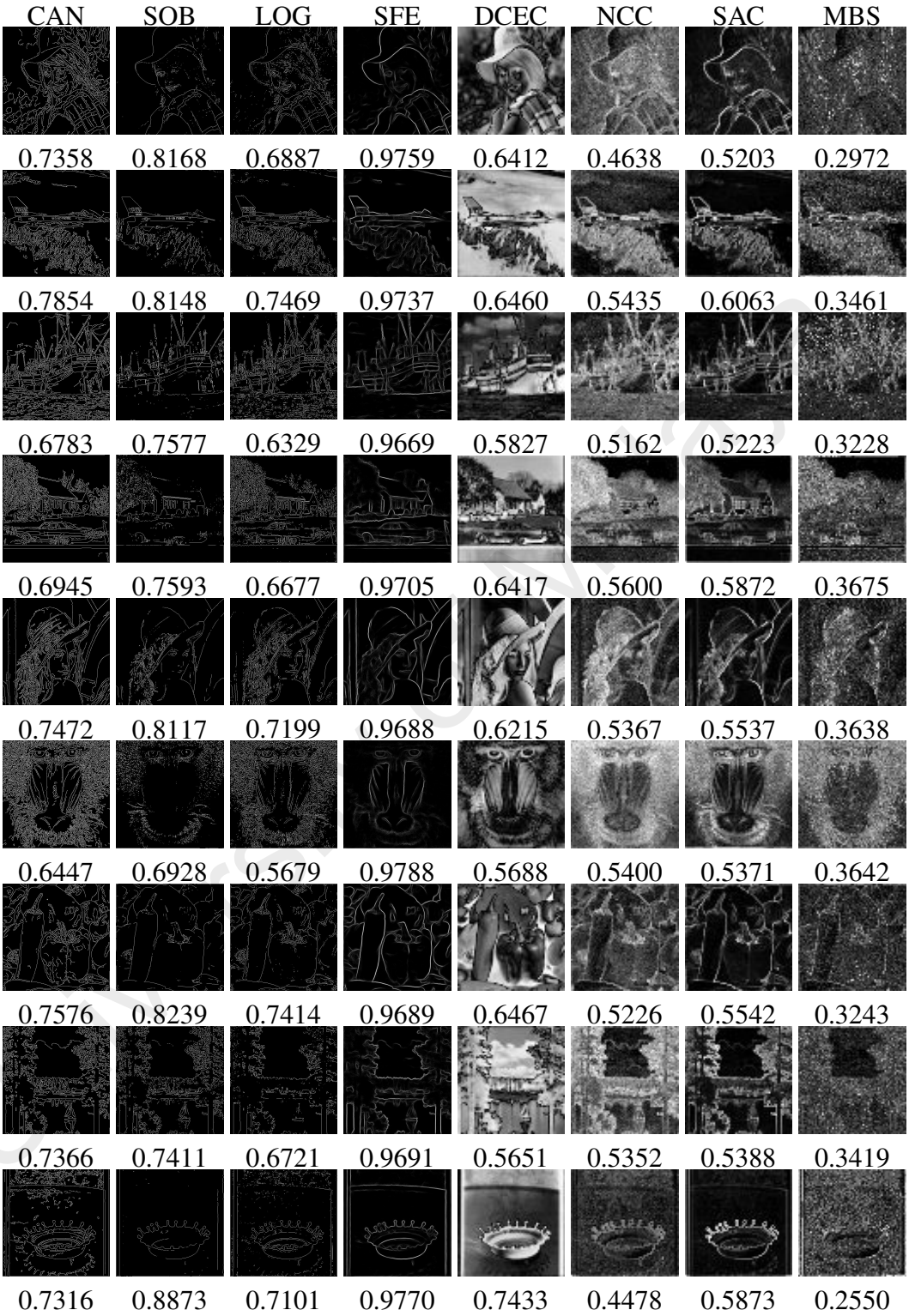
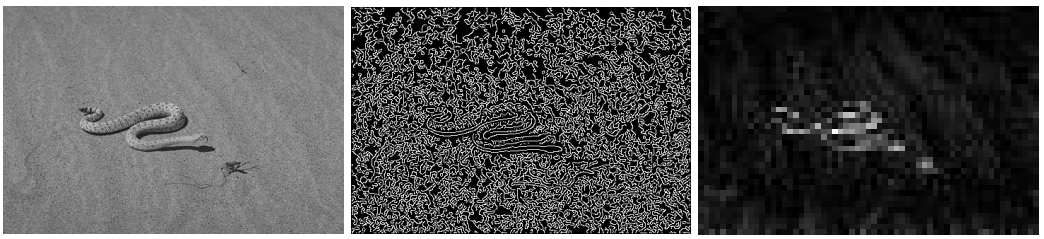


Figure 4.6: Outline images and the corresponding OCA scores for all nine standard test images. From top to bottom: Elaine; F-16; F.B.; House; Lenna; Mandrill; Peppers; S.L., and; Splash



(a) Original (b) CAN (c) DCEC

Figure 4.7: Original image #196073 from the BSDS500 dataset as well as outline images by CAN edge detector and DCEC sketch attack

various complex images, hence, CAN outlines include many edge responses.

Figure 4.7 shows an image from BSDS500 with complex texture and its corresponding outline images generated by CAN and DCEC. The outline image by CAN in Figure 4.7(b) consists of many edge regions, while the outline by DCEC in Figure 4.7(c) shows only the outline of the main object (i.e., snake).

To demonstrate the textural dependency of the proposed assessment metric, the best case (i.e., highest clearness outline image) and the worst case (i.e., lowest clearness outline image) scenarios of BSDS500 are shown in Figure 4.8 and 4.9, respectively. In the best case scenario where the outlines are clear and visible, all outline images have high OCA scores (i.e., > 0.65). On the other hand, in the worst case scenario, except for the outline image generated by SOB and SFE, the outlines are blurred and hardly visible.

Based on the results attained from two datasets, it is concluded that the outline is visible when OCA score ≥ 0.5.

4.5.2 Noise Sensitivity

To verify the sensitivity of the proposed OCA metric, two major distortions are considered, viz., JPEG compression and AGWN, which are the common distortions considered to gauge image clearness/blurring/sharpness assessment metrics (Gu, Zhai, Lin, et al., 2015; Guan et al., 2015; Q. Wu et al., 2015). Note that this chapter did not consider: (a) JPEG2000 because it is not applicable for the sketch attacks considered in this work (i.e.,

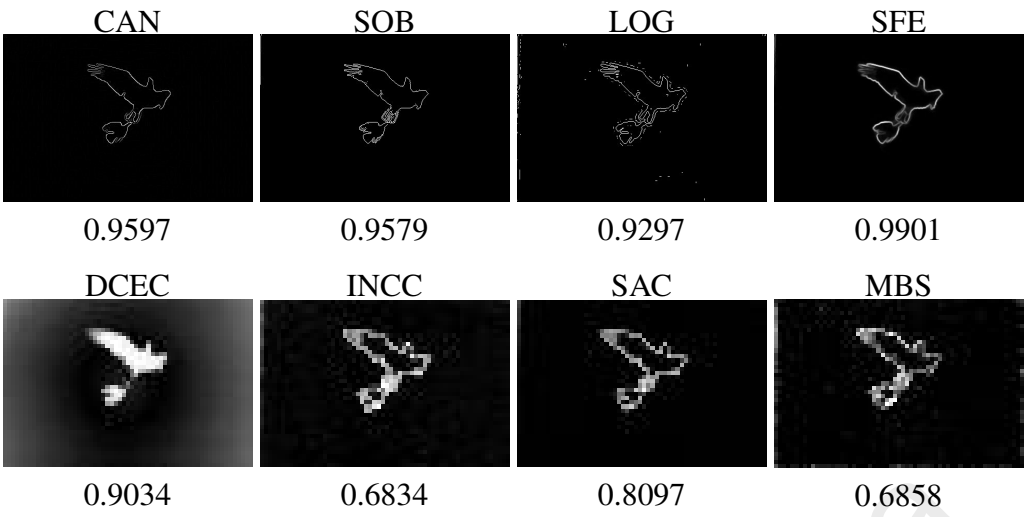


Figure 4.8: Outline images and their OCA scores for image #135069 from BSDS500 (Best case scenario)

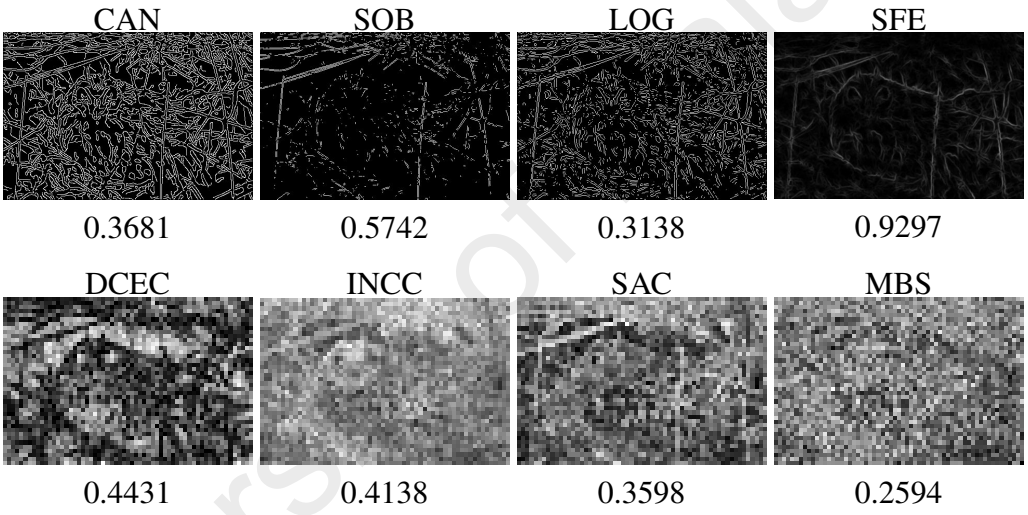


Figure 4.9: Outline images and their OCA scores for image #8143 from BSDS500 (Worst case scenario)

DCEC, INCC, SAC and MBS), and; (b) blurring because for the case of outline image, it is irrelevant to consider the degree of sharpness (or blurring) of an outline image as long as the outline is visible.

Specifically, for JPEG compression, the images in the BSDS500 dataset are encoded into JPEG and then decoded for evaluation. Here, QF is set at 5, 10, 15, $\cdots$, 95. To generate outline images, CAN (Canny, 1986), SOB (Gonzalez & Woods, 2006), LOG (Gonzalez & Woods, 2006) and SFE (Dollar & Zitnick, 2013) consider the pixel values in the fully decoded images, while sketch attacks, DCEC, INCC, SAC and MBS, consider

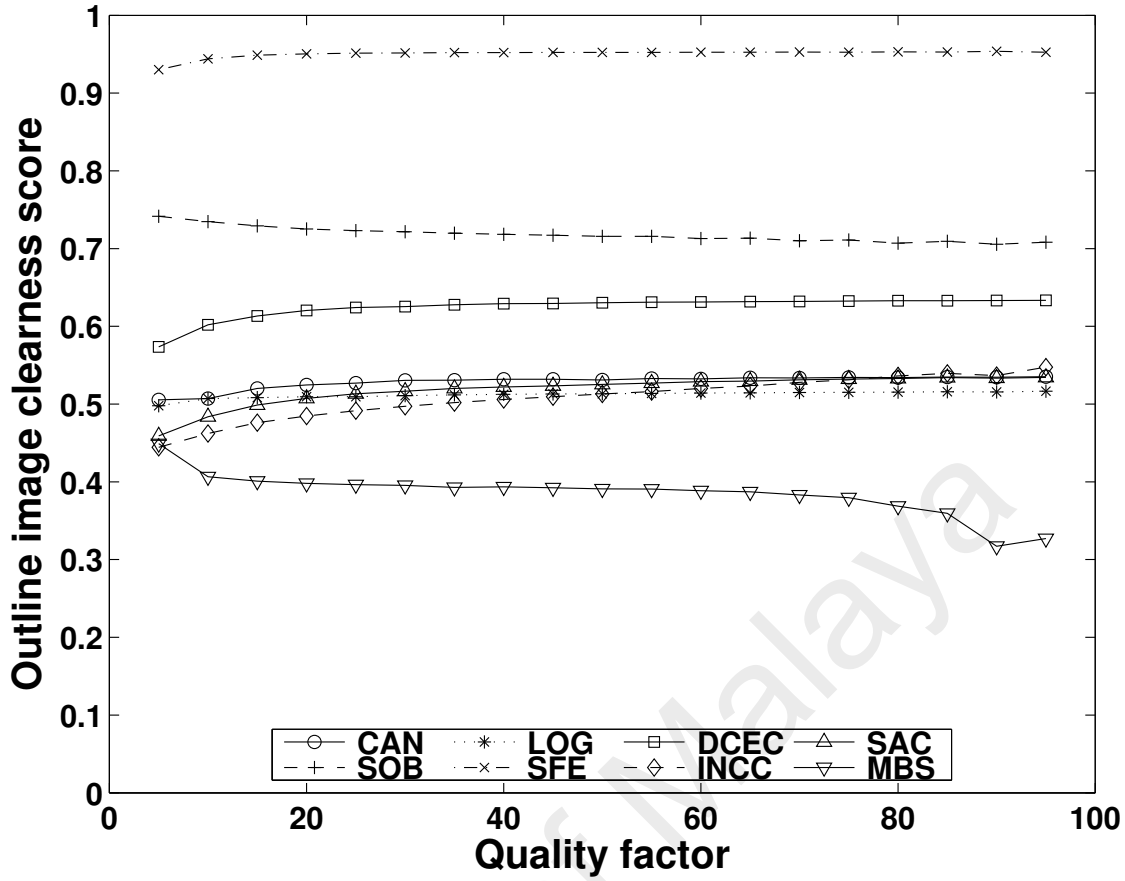


Figure 4.10: OCA scores of BSDS500 dataset for QF $\in [5, 95]$

the partially decoded JPEG bitstream. The graph of OCA scores against QF are shown in Figure 4.10. In general, the proposed OCA score increases steadily as QF increases for CAN, LOG, SFE, DCEC, INCC, and SAC. However, the OCA scores for SOB and MBS decrease as QF increases. It is because the outline by MBS is generated based on the relative number of bits allocated to the blocks, but the number of bits spent on coding insignificant information (as opposed to the actual outline information) increases as QF increases. In addition, SOB uses only simple first-order derivative filters, which is sensitive to noise, and hence SOB fails to differentiate between the increased redundant information and outline information. Similar trends are observed for the standard test image and the results are omitted here.

Next, the OCA score is considered for the images in BSDS500 after applying AGWN with ten standard deviations, viz., $\sigma_w \in \{0, 0.012, 0.024, 0.048, 0.096, 0.125, 0.25, 0.5, 1,$

1.5}. Figure 4.11 shows the graph of OCA score against standard deviation in AGWN. It is observed that the OCA scores drop significantly when σ_w of AGWN increases. Hence, the proposed OCA metric is shown to be sensitive to AGWN. Similar trends are observed for the standard test image and the results are omitted here.

Therefore, the proposed OCA metric is sensitive to very low or very high compression (quality) factors as well as AGWN. It is because these two factors significantly affect the structural properties of any image.

In addition to above discussions, there are some limitations in this work. Although the ideal outline images are provided in some datasets, they are not always available. Therefore, in this chapter, it is assumed that the ideal outline image has only plain regions and perceptible gradients. In other words, a formal definition of outline images is inexistent. This may complicate the assessment of outline images, which is one of the challenges in OCA. The proposed OCA metric may not be able to assess low-resolution outline images, especially, sketch images, which are smaller in terms of image dimension when compared to the input image. Specifically, the subsampling process in a multi-scale analysis may produce an image that is too small, which is not suitable to describe the outline information. Investing into handling low-resolution outline image will be done as future work.

4.5.3 Comparison with Standard Image Quality Assessment

To further discuss the proposed OCA metric, in this section, the comparison with the widely utilized metrics for image quality assessment metrics is discussed. In the literature, there are two major standard image quality metrics, namely: Pixel Signal-to-Noise Ratio (PSNR), which measures the difference of pixel values between a target image and a reference image; and SSIM, which measures the structural similarity of local parts between a target image and a reference image. Although both PSNR and SSIM indicate the

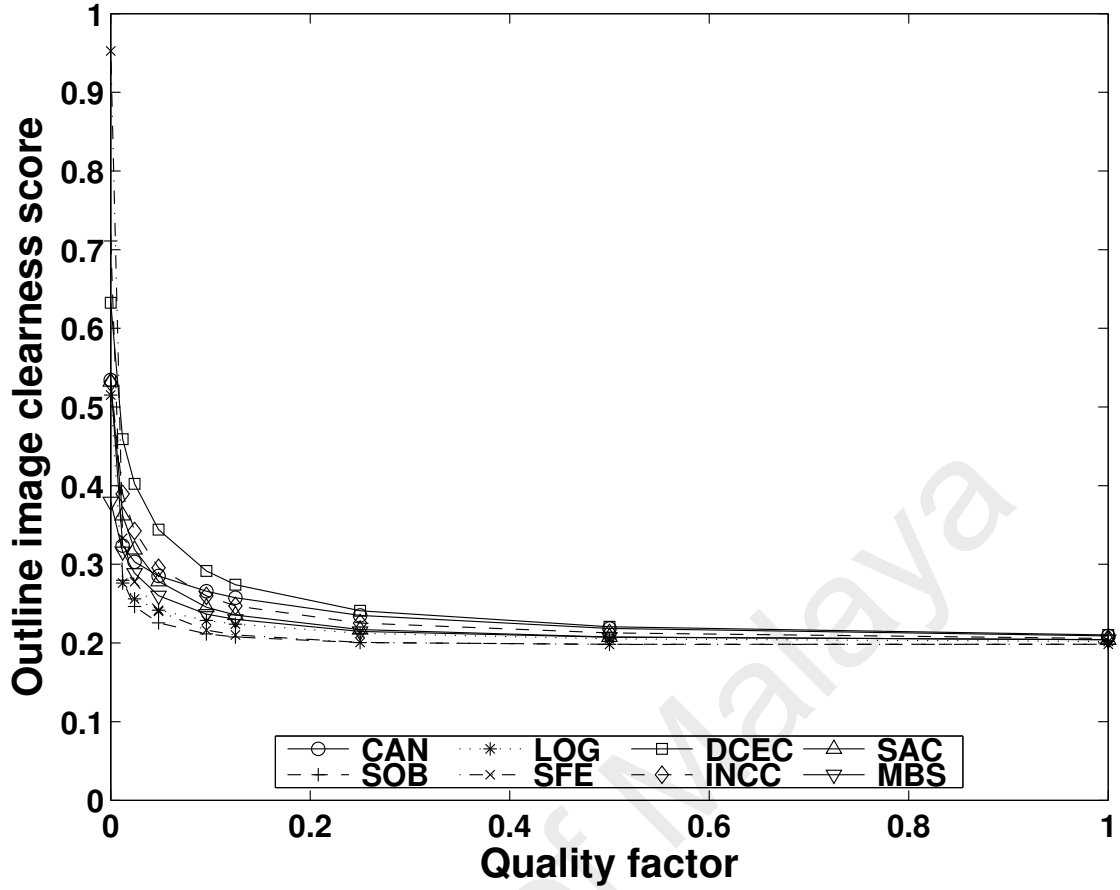


Figure 4.11: OCA scores of BSDS500 dataset for various AGWN (QF = 75)

practical and objective quality of an image, they require a reference image. In addition, they are designed for natural image. On the other hand, the proposed OCA can evaluate sketch image without any reference image, and the OCA scores are found to match our intuitive observation. Therefore, the proposed OCA metric is an appropriate metric for evaluating outline/sketch image.

4.6 Summary

In this Chapter, a no-reference objective metric is proposed to assess the clearness of outline images obtained from edge detectors or sketch attacks. Specifically, the image entropy is exploited to quantify information in an outline image and SSIM is modified into a feasible no-reference clearness metric. The product of these two scores is put forward as an OCA metric. Results suggest that the proposed OCA metric can capture

the outline clearness against compression and it is found to be sensitive to AGWN.

CHAPTER 5: FORMAT-COMPLIANT SELECTIVE ENCRYPTION FRAMEWORK FOR BLOCK-TRANSFORM COMPRESSED IMAGE

5.1 Overview

In this chapter, a general framework is put forward to realize format-compliant selective encryption without bitstream expansion for block-transform compressed images. First, it is discussed that some representative sketch attacks and the newly proposed sketch attack are viable to conventional encryption methods. The proposed sketch attack is utilized to design a format-compliant selective encryption operation. Then, the proposed operation and four encryption operations are augmented to form a complete format-compliant selective encryption method for JPEG compressed image. This chapter ends with the discussion of the distortion, comparisons with conventional methods, and cryptanalysis, to evaluate the performance of the proposed encryption framework.

5.2 Introduction

As aforementioned in Section 2.3, the demand increment of Megapixel (MP) images are elevating the significance of practical security tools for compressed images. In the literature, there are various approaches (Parvin et al., 2014; Subramanyam & Emmanuel, 2014; Tang et al., 2015) to encrypt raw uncompressed images, which can be classified into two classes: Class (a) encryption then compression (Johnson et al., 2004), and; Class (b) compression then encryption.

The latest encryption method (Zhou et al., 2014) of Class (a) achieves high compression efficiency, however, the output is not readily applicable to the existing standards such as JPEG, which is the most widely deployed compression standard. In other words, the method causes the same problem when the standard encryption methods (e.g., DES, AES, or RSA) are applied to the image to encrypt, where the output is non-format compliant.

Hence, format-compliant selective encryption methods are considered in this chapter.

In the literature, various format-compliant selective encryption methods (W. Li & Yuan, 2007; Takayama et al., 2006; Wong & Tanaka, 2010; M. Zhang & Tong, 2014) have been proposed to handle contents stored in the respective formats. However, the conventional format-compliant selective encryption methods for compressed images cause severe bitstream size overhead (W. Li & Yuan, 2007; Wong & Tanaka, 2010; M. Zhang & Tong, 2014), because they improperly treat the components in contents. In addition, the existing format-compliant selective encryption methods can not withstand to the sketch attack (W. Li & Yuan, 2007).

Therefore, this chapter aims to achieve the following: (a) recommending basic requirements in the design of format-compliant selective encryption method for block-transform compressed images; (b) classifying transformed blocks into some groups to make some regions based on the outline coarsely deduced directly from the non-DC coefficients (i.e., AC coefficients in JPEG compression form) to handle the DC coefficients, and; (c) proposing a fully operational format-compliant selective encryption method for JPEG compressed images.

5.3 Requirement for Encryption Framework

Recall that sketch image in Chapter 3 can represent outline image of the input image/video frame, e.g., Figure 5.1(a) shows the DCEC image. Therefore, it is concluded that a viable format-compliant selective encryption framework for block-transform compressed image should take into account the following four requirements: (a) minimum bitstream size overhead; (b) manipulation of the position or value of DC coefficients; (c) manipulation of the number of nonzero AC coefficients in each block of the output (encrypted) image, and; (d) Alternation of the energy of AC coefficients in each output block.

It is found that the conventional methods (W. Li & Yuan, 2007; Takayama et al.,

2006; Wong & Tanaka, 2010; M. Zhang & Tong, 2014) satisfy the requirements (b-d), i.e., withstand against sketch attacks using DC and AC coefficients at the expense of requirement (a). On the other hand, the method (Niu et al., 2008) satisfies the requirement (a), however, it fails the requirements of (b-d).

5.4 Proposed Sketch Attack for Grouping Blocks

The requirement of the bitstream size overhead minimization in Section 5.3 is a challenging task, because it is empirically found that handling DCT coefficient causes bitstream size overhead due to entropy coding. To avoid the overhead, the operations of grouping similar magnitude DC coefficients and handling them are considered in this chapter. To group DC coefficients, a sketch image by a novel sketch attack, namely Energy of AC Coefficient Attack (EAC), is considered, because the sketch image empirically gives better results in handling DC coefficients described in Section 5.5.1. Without loss of generality, let g denote a grayscale image.

The EAC image ϕ_E is defined as follows:

$$\phi_F(i,j) = \text{round}\left(255 \times \frac{e(i,j)}{\bar{e}}\right), \tag{5.1}$$

where

$$e(i,j) = \left(\sum_{u=1}^{8}\sum_{v=1}^{8} |\mathbb{F}_{u,v}(i,j)|\right) - |\mathbb{F}_{1,1}(i,j)|, \tag{5.2}$$

where $e(i,j)$ captures the strength of the edges/textures in the (i,j)-th 8×8 block, $|\mathbb{F}_{u,v}(i,j)|$ denotes the absolute of $\mathbb{F}_{u,v}(i,j)$ and

$$\bar{e} = \frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N} e(i,j). \tag{5.3}$$

Figure 5.1(b) shows the sketch image ϕ_E. The strength of the edges/textures, if any,

(a) DCEC (b) EAC

Figure 5.1: Sketch images by applying the sketch attacks: (a) DCEC, and; (b) EAC to a JPEG compressed image

founded in the block, can be represented by the energy of AC coefficients, where higher AC energy suggests stronger edges/textures, and vice versa.

5.5 Proposed Encryption Framework

In this section, an encryption framework is proposed to encrypt images while satisfying the four requirements. In the following sub-sections, the operation of rearranging the DC coefficients is detailed, and then additional operations for format-compliant selective encryption handling non-DC coefficients (i.e., AC) are also detailed to form a complete format-compliant selective encryption method for JPEG compressed images. In addition, the format-compliant selective decryption is also detailed.

5.5.1 Rearranging DC Coefficients (RDC)

In this section, a novel approach, namely RDC operation, is put forward to efficiently encrypt the DC coefficients in a block-transform compressed image. The conventional operations can be classified into either (a) shuffle the DC coefficients (W. Zeng & Lei, 2003); or (b) map the errors of DC coefficients to other different values in the same category (Niu et al., 2008). However, methods in the class (a) result in bitstream size overhead, while operations in class (b) are vulnerable to DCEC attack. Hence, to avoid

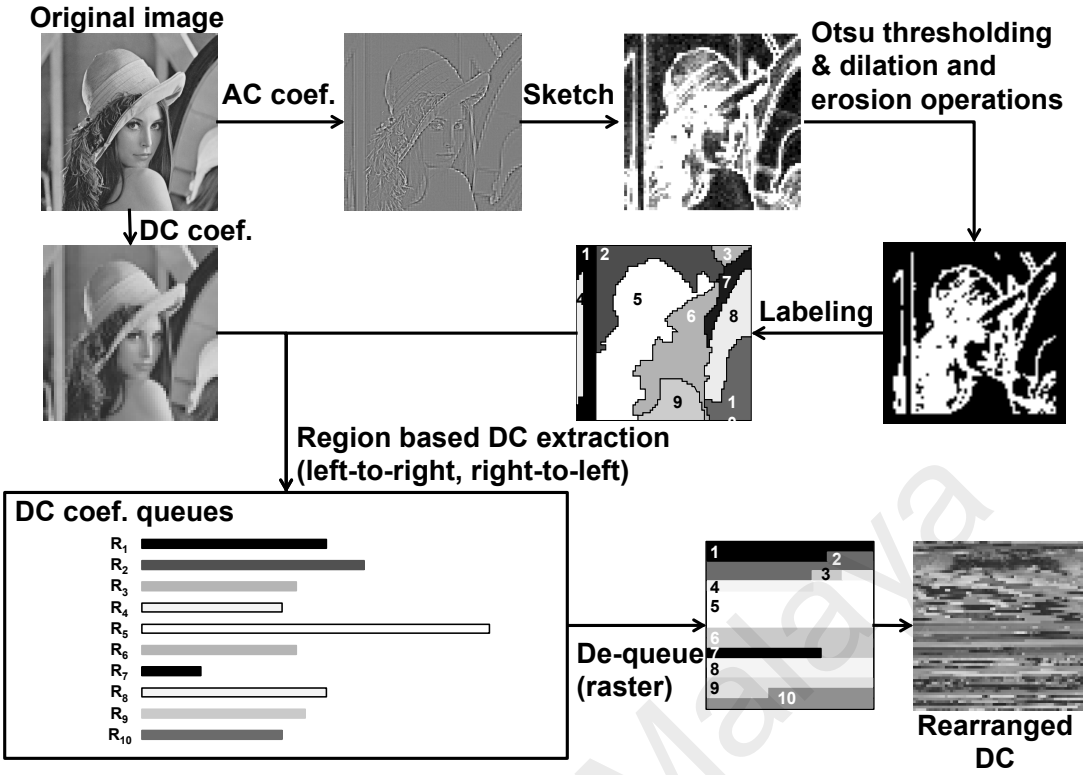


Figure 5.2: Process flow of operations in RDC

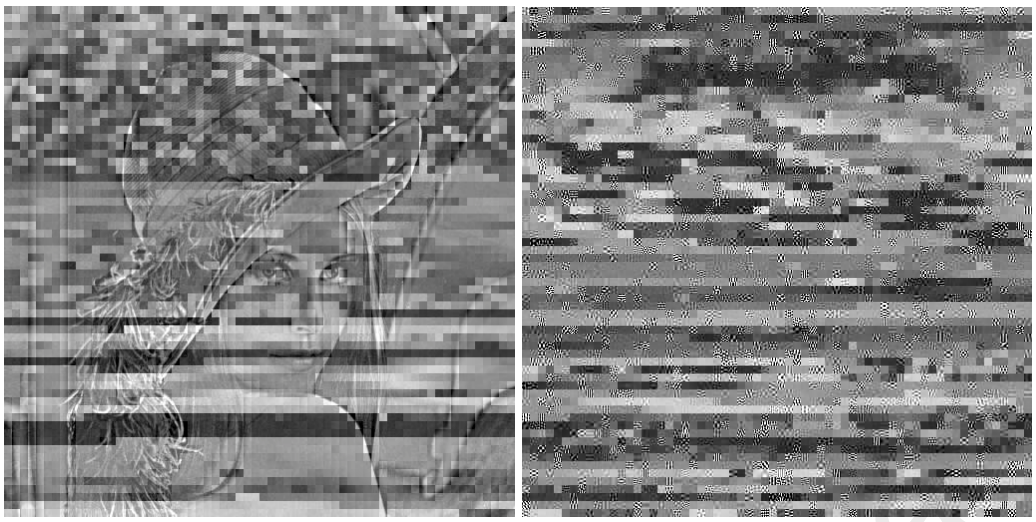
these drawbacks of the conventional operations, the proposed operation is designed to rearrange the DC coefficients, where distortion of the image quality and minimization of bitstream size overhead are achieved simultaneously.

In the ideal case, when the DC coefficients are sorted in non-decreasing (or non-increasing) order prior to the prediction process, the compression gain is maximized. However, the sorting causes the location information loss of each DC coefficient and the overhead for sorting the original locations is incurred. The overhead is far higher than the compression gained by the sorting. Therefore, a sub-optimal approach is adopted in the proposed operation without invoking the sorting process. The considered observations in the idea of designing sub-optimal approach are as follows: (a) the intensity values of pixels within an object (or a connected region) are relatively similar values (i.e., small prediction error), and; (b) a DC coefficient value is the average intensity value of its corresponding block.

Figure 5.2 shows the process flow of operations in Rearranging DC Coefficients

(RDC). First, EAC is applied to the original (plaintext) image for obtaining its sketch image, where the information of the number of nonzero AC coefficients in each block is extracted by Equation 5.1. This operation is viable because the AC coefficients in each block are still intact (i.e., not handled yet). Note that other sketch attacks in Chapter 3 are able to be utilized instead of EAC, but the same sketch attack must be used for during format-compliant selective encryption and during reconstruction. Then, Otsu's thresholding method with the default parameter value of 0.5 (Otsu, 1979) is applied to the obtained sketch image for generating a binary image. To separate weakly connected lines and remove isolated points in the binary image, morphological operations, viz., erosion followed by dilation of window size 3×3 pixels (i.e., opening operation), are applied to the binary image. Then, a labeling operation (Robert & Linda, 1992) with 4-connectivity is applied to the processed binary image. The DC coefficients of each labeled region are extracted in the horizontal manner (i.e., left-to-right on odd row, right-to-left on even row). To form a further processed queue, the extracted DC coefficients are sequentially added to the queue. To avoid loss of generality, the regions are handled in ascending order with respect to their labels. Last, to form rearranged DC coefficients, the queued DC coefficients are sequentially extracted and added to the positions designed for the DC coefficients in raster order.

Since this rearranging DC coefficient position doesn't change DC coefficient values, a format-compliant selective encryption withstanding significant bitstream size overhead can be expected. However, the processed image by RDC still has a recognizable outline of the original image as shown in Figure 5.3(a). The result is expected the outcome because the non-DC (i.e., AC) coefficients carry the textural/edge information of the original image and they are unaltered. Therefore, the JPEG encoder is modified to include encryption operations in the following sub-section for encryption the DC and AC coefficients during encoding. Note that the proposed format-compliant selective encryption is



(a) Processed DC only (b) Encrypted image

Figure 5.3: Example of processed JPEG images

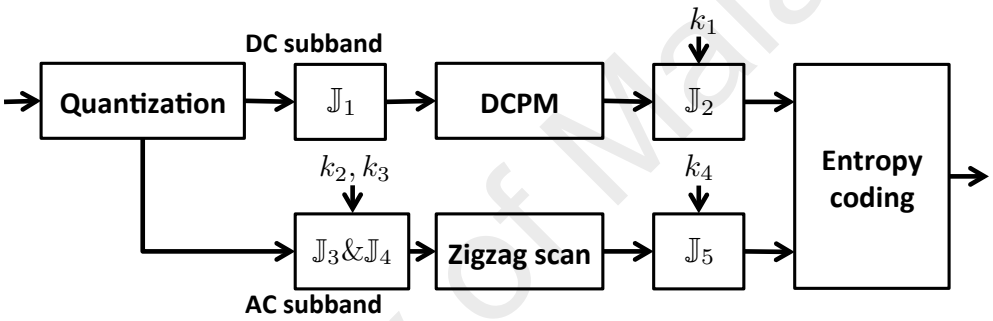


Figure 5.4: Modified JPEG encoder

directly applicable to JPEG compressed images.

5.5.2 Encryption Method for JPEG

To perform entire format-compliant selective encryption operations, the JPEG encoder is modified as shown in Figure 5.4. The modified encoder includes five format-compliant selective encryption operations, viz., RDC (denoted as $\mathbb{J}_1$) and a conventional operation for the DC (denoted as $\mathbb{J}_2$) and three conventional operations for the AC coefficients (denoted as $\mathbb{J}_3$, $\mathbb{J}_4$, and $\mathbb{J}_5$), are consolidated for format-compliant selective encryption method:

$\mathbb{J}_1$ RDC: The DC coefficients of an input image are rearranged in grouped regions based on the labeled regions as described in Section 5.5.1. Note that the region

labels can be shuffled further to withstand unauthorized reviewing or reconstruction of the original image.

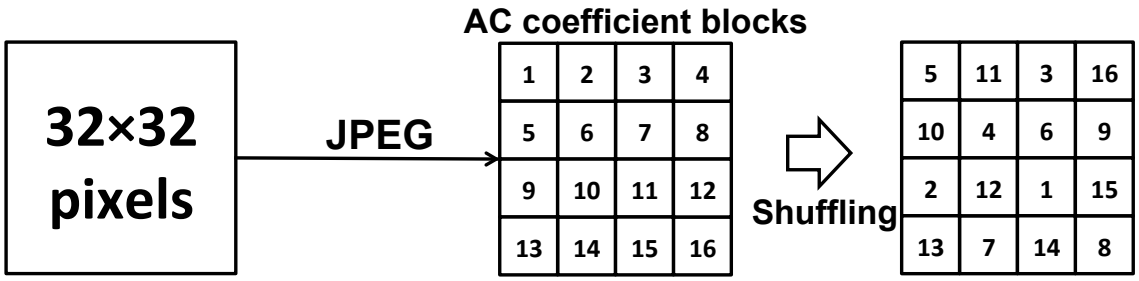
$\mathbb{J}_2$ DC Error Category Mapping: A bijective mapping method is applied to the each predicted error to obtain a value from the same category as in (Niu et al., 2008). The mapping method maintains the same sign of the predicted error, because randomizing sign of the predicted error may cause underflow/overflow problem in a general decoder. To generate encryption map for the predicted error, a secret key k_1 is considered.

$\mathbb{J}_3$ AC Sign Randomization: Each nonzero AC coefficient sign (i.e., $-1/1$) is multiplied with a pseudo-random sequence of -1's and 1's (Lian et al., 2007) using a secret key k_2.

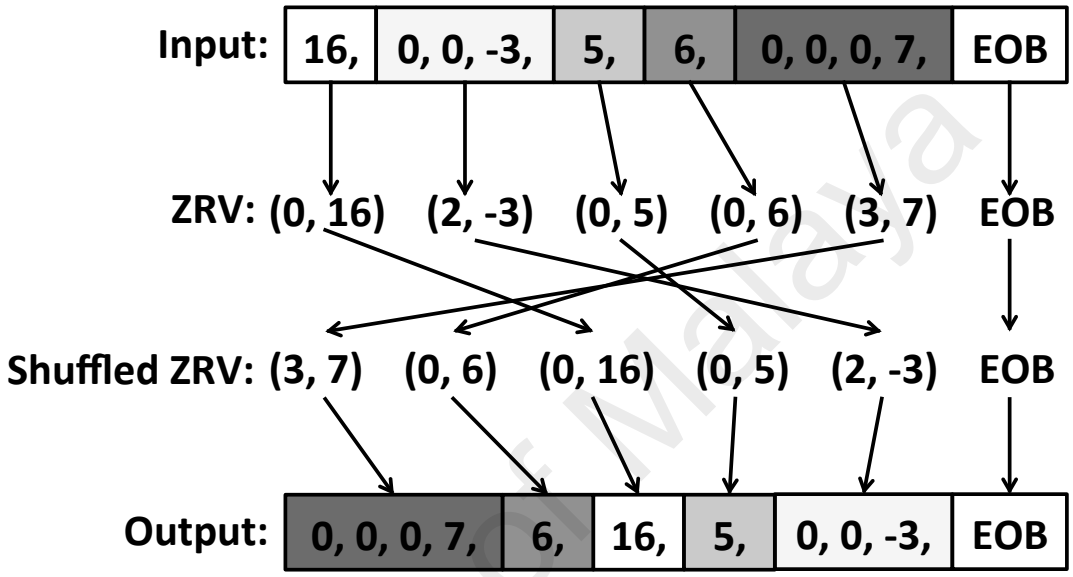
$\mathbb{J}_4$ AC Block Shuffling: The blocks consisting only modified AC coefficients are then globally shuffled using Equation (2.4) (Takayama et al., 2006) with the random permutation map generated with a secret key k_3. For example, $\mathbb{B}'_{AC}(1,2) = \mathbb{B}_{AC}(1,3)$, and so forth.

$\mathbb{J}_5$ AC ZRV Pair Shuffling: The ZRV pairs within each block are shuffled as in Reference (Takayama et al., 2006) using a secret key k_4.

An example image (Lenna) encrypted with the proposed encryption method is shown in Figure 5.3(b). It is found that the outline of Lenna is a completely unintelligible image, hence the noise caused by both processing the DC and AC coefficients is sufficient to distort images. Furthermore, the distortion strength level can be intensified by replacing the entries of the quantization table with larger numbers for changing the pixel values (to some very small/large values).



(a) Block shuffling ($\mathbb{J}_4$)



(b) ZRV pairs shuffling ($\mathbb{J}_5$)

Figure 5.5: ZRV pairs shuffling and block shuffling

To facilitate the understanding of the proposed format-compliant selective encryption operations, Figure 5.5 illustrates the operations of AC block shuffling and ZRV shuffling. Specifically, Figure 5.5(a) shows the results of AC block shuffling (i.e., $\mathbb{J}_4$) for AC blocks of size 32×32 pixels, where the numbers represent the indices for the AC blocks. On the other hand, Figure 5.5(b) shows the output of shuffling ZRV pairs (i.e., $\mathbb{J}_5$) using 16, 0, 0, -3, 5, 6, 0, 0, 0, and -7 as the representative sequence of AC coefficients.

5.6 Decrypting Method for JPEG

In this section, the decryption operations to reconstruct the original image from its encrypted counterpart are described. Our modified JPEG decoder consisting five operations is illustrated in Figure 5.6. The five operations, namely the two decryption operations

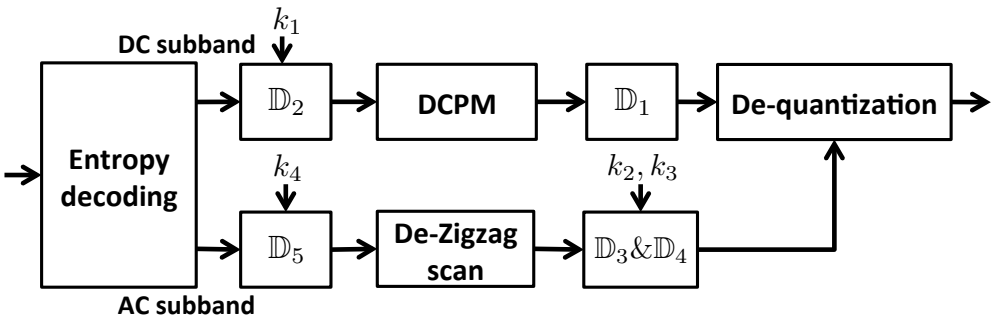


Figure 5.6: Modified JPEG decoder

for the DC coefficients (denoted as $\mathbb{D}_1$ and $\mathbb{D}_2$) and the three format-compliant selective decryption operations for the AC coefficients (denoted as $\mathbb{D}_3$, $\mathbb{D}_4$ and $\mathbb{D}_5$), are consolidated for reversing the processes performed during format-compliant selective encryption method:

$\mathbb{D}_1$ Inverse RDC: The DC coefficients of an encrypted image by the proposed encryption method are rearranged in the correct (original) locations using the connected regions of the binary image deduced from the sketch image of the original image. The inverse RDC will be detailed in the last paragraph of this section.

$\mathbb{D}_2$ DC Prediction Error Decryption: To obtain the correct (original) value, each the encrypted DC prediction error is remapped by using the same key k_1 as in the modified encoder for format-compliant selective encryption.

$\mathbb{D}_3$ AC Sign Reconstruction: To reconstruct each correct (original) sign, the correct pseudo-randomly generated a sequence of -1's and 1's (Lian et al., 2007) by using k_2 is multiplied to the sign of each decoded nonzero AC coefficient.

$\mathbb{D}_4$ AC Block Reconstruction: The correct position blocks (consisting of AC components only) are reconstructed as in (Takayama et al., 2006) with the correct index generated by k_3.

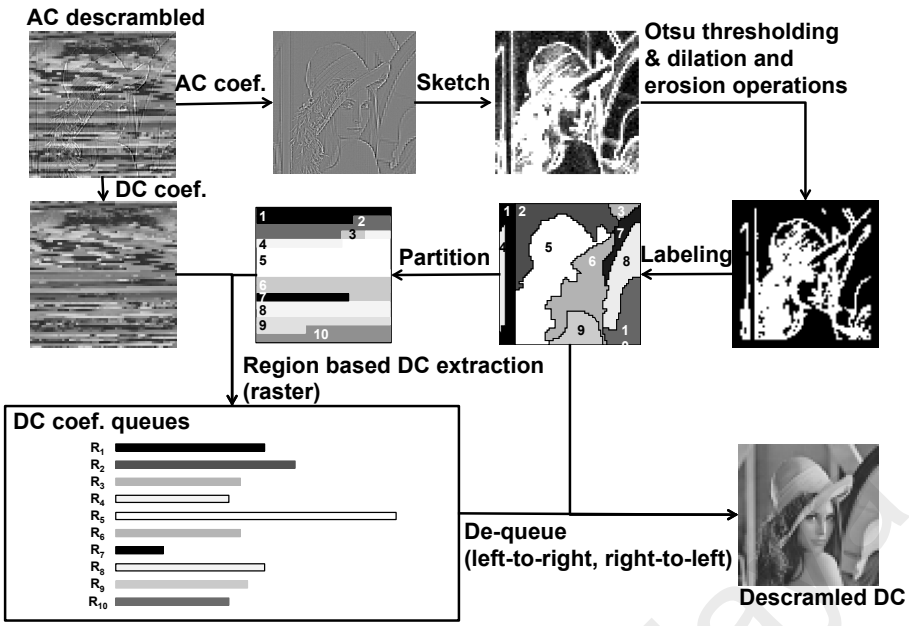


Figure 5.7: Process flow of inverse RDC coefficients

$\mathbb{D}_5$ AC ZRV Pair Decryption: ZRV pairs within a block are decrypted as in Reference (Takayama et al., 2006) using k_4.

Unlike the format-compliant selective encryption method, the AC coefficients are processed first, and the reconstructed blocks of AC coefficients are utilized to obtain regions deduced from the sketch image of the original image for restoring the positions of the DC coefficients. Specifically, the decryption process follows the flow order of D_5, D_4, D_3, D_2, and then D_1.

Figure 5.7 shows the inverse RDC operation flow after completely reconstructing AC coefficients, where the connected regions deduced from the sketch image are utilized to restore the DC coefficients. Specifically, first the AC coefficients are fully reconstructed (i.e., operation D_5, D_4, and D_3), the sketch image of the image is obtained by using the same sketch attack (i.e., either INCC or EAC) as in the modified encoder. Then, the regions in the binary image deduced from the sketch image are grouped and labeled in the same manner as in the modified encoder. The DC coefficients of each region are sequentially reserved in a fixed length queue. Specifically, starting from label index 1, the DC coefficients are extracted in the raster scan order and put in the queue. When a queue

becomes full with DC coefficients, the following DC coefficients are put in the sequential queues with label indices 2, 3, and so forth until DC coefficients are processed. Finally, by using the order of left-to-right on odd rows and right-to-left on even rows, the queues of DC coefficients are de-queued to fill each labeled region.

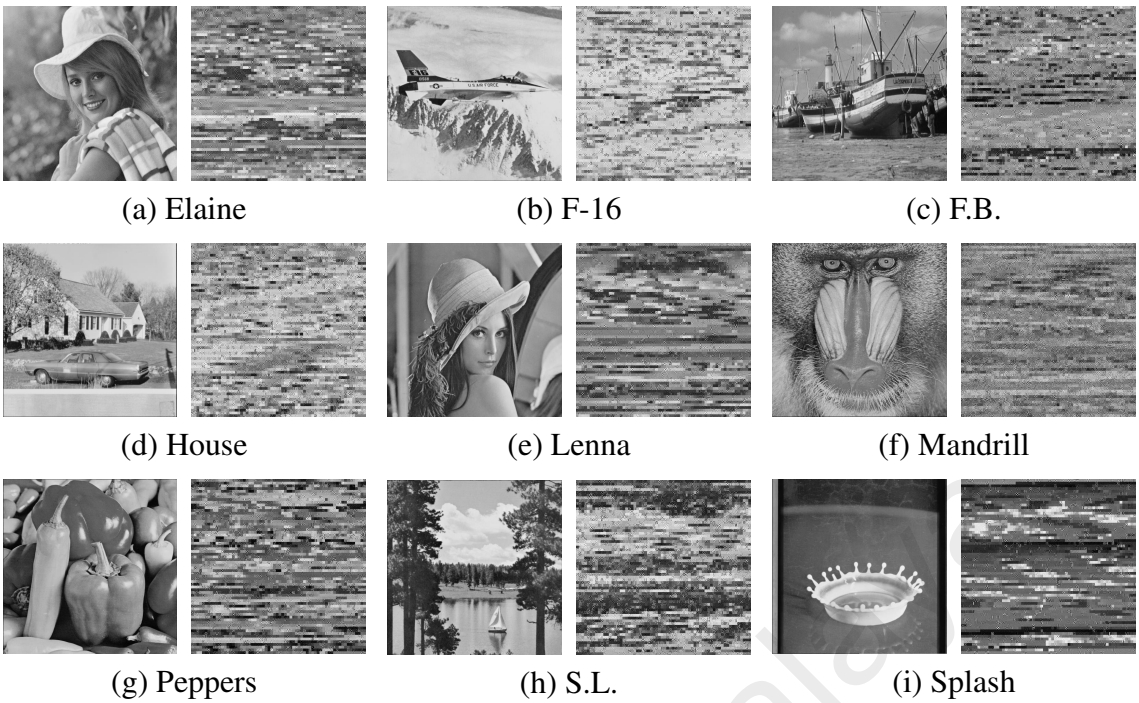
5.7 Experiment Results and Discussions

To verify the performance of the proposed format-compliant selective encryption, the method is implemented using the reference software (Group's, 2012) and two datasets, i.e., nine standard test images (USC-SIPI, 2014) and the *SIMPLIcity* image dataset (Z. J. Wang et al., 2001) are considered, respectively, where the nine standard test images are in grayscale with the size of 512×512 pixels each (i.e., Elaine, F-16, F.B. , House, Lenna, Mandrill, Peppers, S.L., and Splash) and the SIMPLIcity image dataset consists of 1000 images (i.e., mixture of building, human, coast, train, animal, flower, etc) with the size of 256×384 pixels. In addition, to generate pseudo-random numbers, *Mersenne Twister* (Matsumoto & Nishimura, 1998) is utilized with a 32-bit seed.

The experiment results generated by the proposed JPEG format-compliant encryption method are discussed in the following sub-sections. To avoid loss of generality, the parameter of QF is fixed at 75 and all QT values are set to 255. First, the results in terms of distortion are presented in Section 5.7.1. Next, the comparison results among the proposed and conventional encryption methods are discussed in terms of robustness against sketch attacks and bitstream size overhead in Section 5.7.2. Then, security analysis is verified in Section 5.7.3. Last, the limitation of the proposed method is summarized in Section 5.7.4.

5.7.1 Distortion

The nine original images and the corresponding output images generated by the proposed format-compliant selective encryption method are shown in Figure 5.8, respectively. It



(a) Elaine (b) F-16 (c) F.B.

(d) House (e) Lenna (f) Mandrill

(g) Peppers (h) S.L. (i) Splash

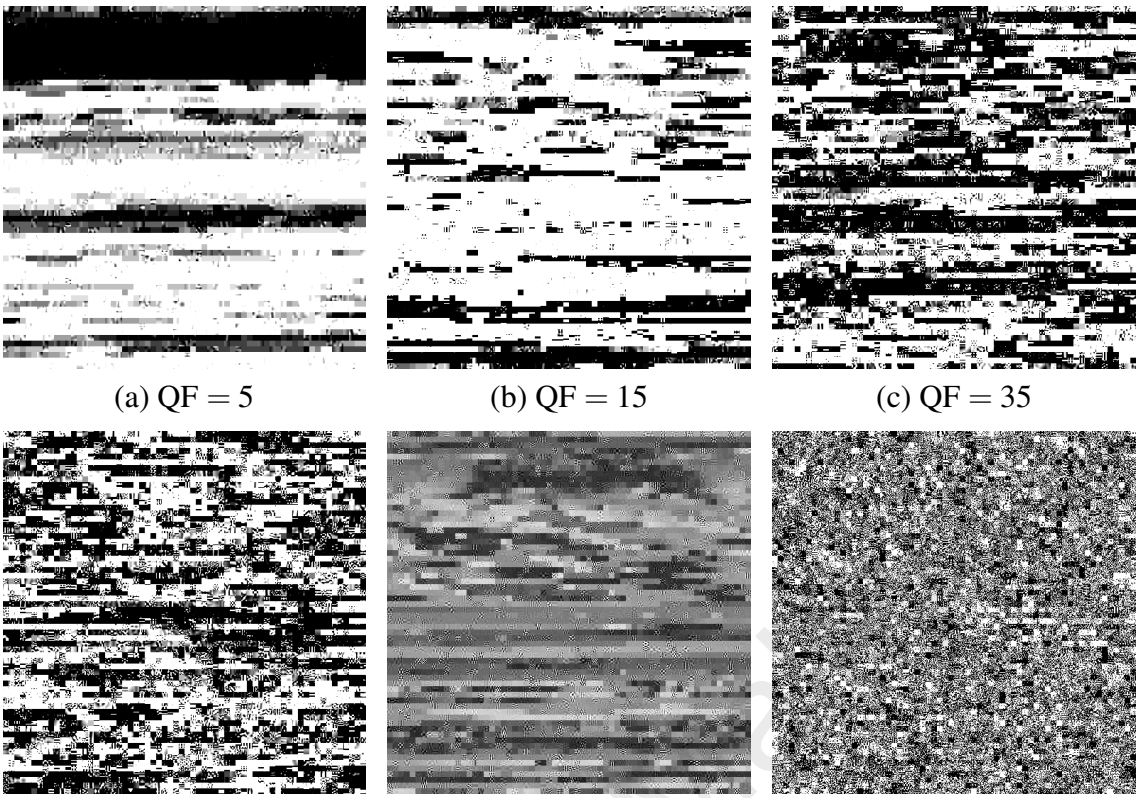
Figure 5.8: Original images and outputs of the proposed encryption method for various JPEG compressed images

Table 5.1: SSIM and PSNR (dB) of output images generated by the proposed method

Image	SSIM	PSNR (dB)
Elaine	0.1100	6.1808
F-16	0.1205	5.4378
F.B.	0.0624	6.1711
House	0.0668	5.8528
Lenna	0.0903	6.1316
Mandrill	0.0227	6.3116
Peppers	0.0866	6.3002
S.L.	0.0495	5.9968
Splash	0.1882	5.6295
SIMPLIcity	0.0779	5.8275

is clearly found that the proposed method transforms each image into an unintelligible image regardless of its original texture/edge. To quantify the obtained distortion level, the SSIM (Z. Wang et al., 2004) and PSNR Decibel (dB) values of the encrypted images are recorded in the second and third columns of Table 5.1, respectively. Results suggest that SSIM and PSNR dB values are very low, in other words, the encrypted images by the proposed encryption method are significantly distorted.

To proof of the concept, the encrypted Lenna images using various QF $\in \{5, 15,$



(a) QF = 5 (b) QF = 15 (c) QF = 35

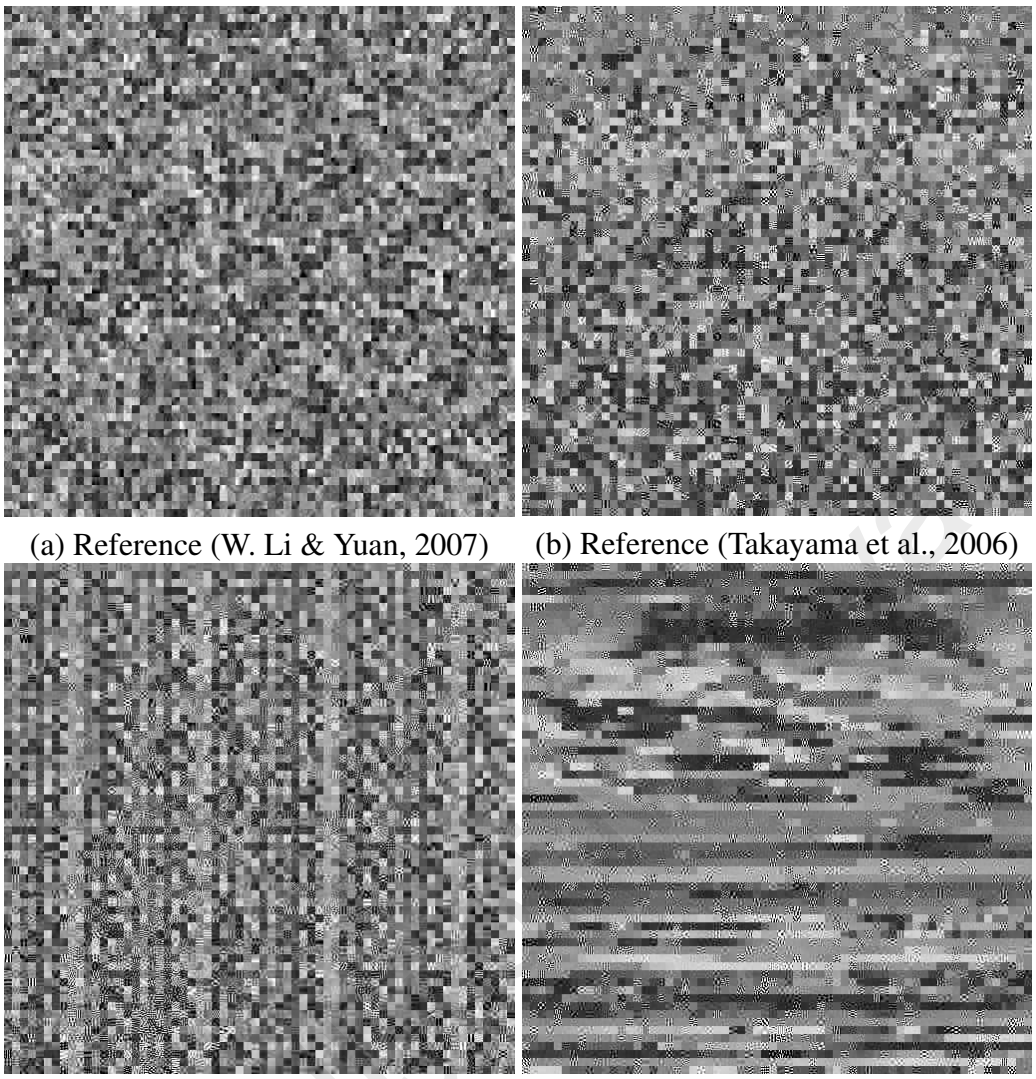
(d) QF = 55 (e) QF = 75 (f) QF = 95

Figure 5.9: Encrypted Lenna images for various JPEG quality factors

35, 55, 75, 95} are shown in Figure 5.9, where QF = 5 and 95 represent the low and high fidelity images, respectively. The encrypted images are unintelligible, hence results suggest that the proposed encryption method is viable to perceptually distort the images compressed with different QFs. It is found that the distortion becomes more stronger when QF increases, because high fidelity image generally has more nonzero coefficients, which directly contribute to the appearance/distortion of the image. Similar results of other test images are also obtained.

5.7.2 Comparisons with Conventional Methods

In this section, the performances (i.e., spatial correlation, sketch attack robustness and bitstream size overhead) of the proposed method is compared with those of the conventional format-compliant selective encryption methods (W. Li & Yuan, 2007; Niu et al., 2008; Takayama et al., 2006; Wong & Tanaka, 2010). As shown in Figure 5.10, the considered methods are viable to distort the compressed image into completely unintel-



(a) Reference (W. Li & Yuan, 2007) (b) Reference (Takayama et al., 2006)

(c) Reference (Wong & Tanaka, 2010) (d) Proposed method

Figure 5.10: Distorted images for the test image Lenna encrypted by all format-compliant selective encryption methods considered (QF = 95)

ligible form, therefore the comparison in terms of image quality is omitted. Here, the Reference (M. Zhang & Tong, 2014) is not considered since the method causes severe bitstream size overhead (i.e., $> 50\%$).

First, spatial correlations between neighboring pixels of the encrypted images are considered. Generally, format-compliant selective encryption methods are designed to conceal the perceptual information of the original image, in other words, the encryption methods destroy the correlation among pixels. Therefore, lower spatial correction in any direction implies better performance of encryption. The spatial correlation value $\omega_{\mathbb{PO}}$ can be obtained as follows:

$$\omega_{\mathbb{PQ}} = \frac{E[(p-\mu_{\mathbb{P}})(o-\mu_{\mathbb{O}})]}{\sigma_{\mathbb{P}}\sigma_{\mathbb{O}}}, \tag{5.4}$$

where E is the expectation function, $\mathbb{O}$ is the set of pixel values in g, $\mu_{\mathbb{O}}$ is the mean pixel value of the input image g, and $\sigma_{\mathbb{O}}$ is the variance of the pixel values in g. $\mathbb{P}$, $\mu_{\mathbb{P}}$ and $\sigma_{\mathbb{P}}$ are defined similarly for the shifted image g_α. Here, g_α assumes the values of *hor* $g_{\text{hor}}(x,y)$, *ver* $g_{\text{ver}}(x,y)$, *dia* $g_{\text{dia}}(x,y)$, denoting the direction of right, bottom and diagonal (i.e., bottom-right), respectively. Specifically, the shifted image g_α is constructed as follows:

$$g_{\text{hor}}(x,y) = g(x,y+1) \tag{5.5}$$

$$g_{\text{ver}}(x,y) = g(x+1,y) \tag{5.6}$$

$$g_{\text{dia}}(x,y) = g(x+1,y+1) \tag{5.7}$$

where $1 \leq x \leq X$ and $1 \leq y \leq Y$.

Table 5.2 shows the spatial correlation values for the horizontal, vertical and diagonal directions, where *Original* denotes that no encryption operations are applied to the input image, namely the output is the original image. Results suggest that the original image yields the highest correlation value in any directions, and the proposed method yields the lowest correlation value among the considered format-compliant selective encryption methods. Therefore, from destroying spatial correlation point of view, the proposed method is able to outperform the considered format-compliant selective encryption methods.

Next, the comparison of the robustness against sketch attack among the considered methods is considered. Only (Niu et al., 2008) is vulnerable to the sketch attacks con-

Table 5.2: Spatial correlation in the horizontal, vertical and diagonal directions (using Lenna compressed at QF = 75 as the test image)

Method	Horizontal	Vertical	Diagonal
Original	0.9737	0.9868	0.9609
Proposed	0.3811	0.3802	0.2697
W. Li & Yuan (2007)	0.6102	0.6495	0.4485
Takayama et al. (2006)	0.8459	0.8590	0.7302
Wong & Tanaka (2010)	0.5710	0.6206	0.3941

sidered, while the rest are able to withstand them (i.e., DCEC, INCC, and EAC). It is because (Niu et al., 2008) changes neither the category nor the position of the DC coefficients, in other words, requirement (a) is satisfied. Therefore, when DCEC is applied to the Lenna image encrypted by (Niu et al., 2008), an output similar to Figure 5.1(a) can be obtained. On the other hand, (W. Li & Yuan, 2007; Takayama et al., 2006; Wong & Tanaka, 2010) and the proposed encryption method are robust against the sketch attacks considered because they satisfy the requirement (b-d).

Last, the bitstream size overhead by encryption is considered, Table 5.3 shows the bitstream size overhead of the encrypted images by the considered methods, where the results for (Niu et al., 2008) is not recorded because they are always near to zero. Results clearly show that the considered conventional methods (W. Li & Yuan, 2007; Takayama et al., 2006; Wong & Tanaka, 2010) suffer from bitstream size overhead. Specifically, Reference (W. Li & Yuan, 2007) induces severe bitstream size overhead, i.e., $\sim +20\%$ of the original image. On the other hand, Reference (Niu et al., 2008) and the proposed method are viable to suppress bitstream size overhead, where their overhead are significantly lower than those of the other considered methods (W. Li & Yuan, 2007; Takayama et al., 2006; Wong & Tanaka, 2010). The overhead of the proposed method and the considered methods (W. Li & Yuan, 2007; Takayama et al., 2006; Wong & Tanaka, 2010) are, on average, $+0.032\%$, $+3.518\%$, $+19.770\%$, and $+11.955\%$, respectively. Note that similar trends are observed for other QF values. Figure 5.11 shows the average bit-

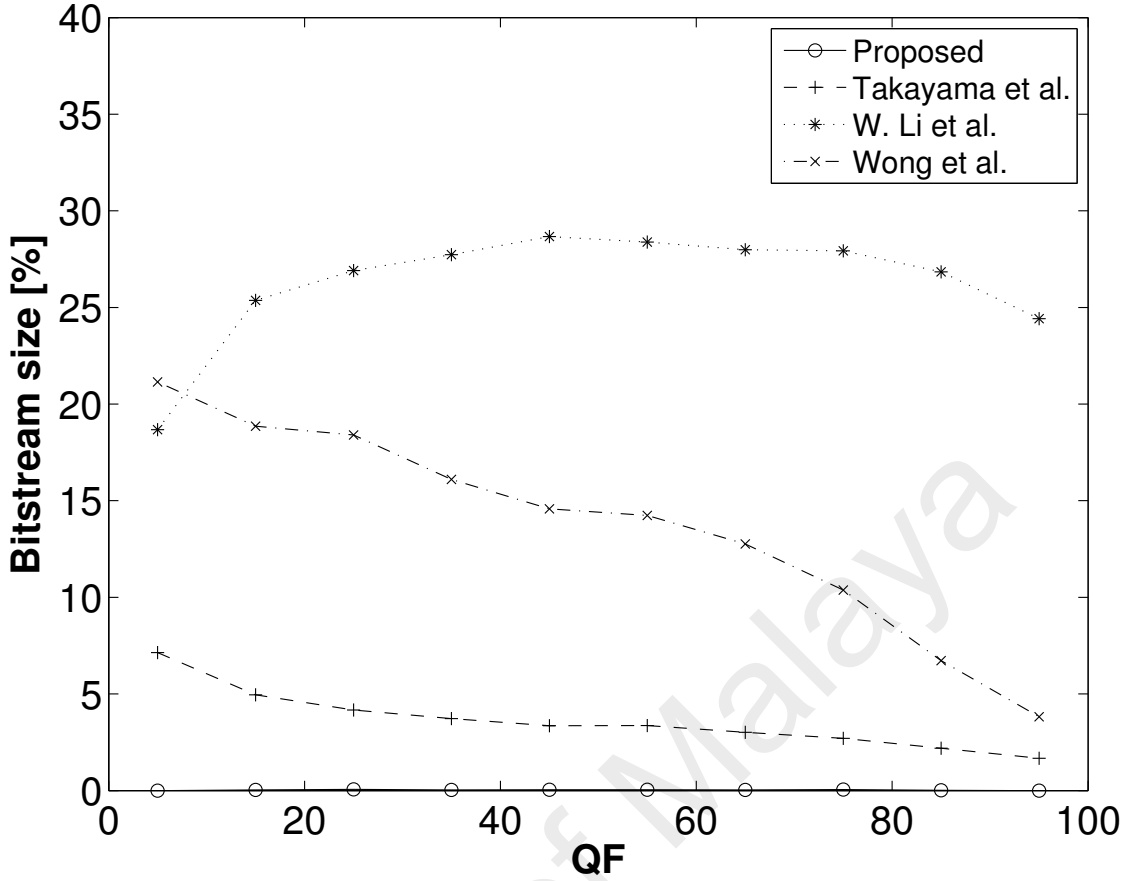


Figure 5.11: Graph of average bitstream size overhead against quality factor for the SIMPLIcity dataset

stream size overhead for various QFs with the SIMPLIcity dataset (Z. J. Wang et al., 2001). It is clearly found that the proposed encryption method stably suppresses the bitstream size overhead (i.e., $+0\%$) for all QFs considered, while the conventional methods cause different results.

Table 5.3: Percentage of bitstream size expansion for JPEG (QF = 75)[%]

Image	Proposed	Takayama et al. (2006)	W. Li & Yuan (2007)	Wong & Tanaka (2010)
Elaine	+0.615	+3.464	+9.258	+11.565
F-16	+0.044	+3.985	+25.661	+12.980
F.B.	+0.141	+2.468	+17.224	+10.572
House	+0.067	+2.719	+23.199	+10.904
Lenna	-0.241	+4.145	+21.445	+14.404
Mandrill	+0.092	+1.459	+10.673	+6.396
Peppers	+0.010	+4.506	+17.836	+14.214
S.L.	-0.144	+3.422	+16.640	+11.054
Splash	-0.274	+5.383	+17.319	+16.694
SIMPLIcity	+0.04	+2.698	+27.931	+10.377
Average	+0.032	+3.518	+19.770	+11.955

Recall that (Niu et al., 2008) is able to suppress the bitstream size overhead, but it is vulnerable to the sketch attack DCEC. On the other hand, methods (W. Li & Yuan, 2007; Takayama et al., 2006; Wong & Tanaka, 2010) can withstand the considered sketch attacks at the expense of bitstream size overhead. Therefore, it is concluded that the proposed method is able to achieve outperformed-balanced performance which is the robustness against sketch attack as well as minimizing bitstream size overhead.

5.7.3 Cryptanalysis

In this sub-section, the robustness of the proposed encryption method against conventional cryptanalysis is analyzed. By attacking, it is assumed that the adversary does not have the keys used during encryption when given only an encrypted image.

In terms of cryptanalysis, a format-compliant selective encryption method is called robust if the number of all trial combinations is infeasible to reconstruct the original image from the encrypted image, which approach is referred to as the plaintext only (i.e., brute force) attack in the literature. To avoid loss of generality, let $n(i, j)$ denote the number of nonzero AC coefficients in the (i, j)-th 8×8 block in a JPEG compressed image and θ denote the number of regions deduced from the sketch image considered in RDC. The total number of possible combinations for each process (i.e., $\tau(\mathbb{J}_1)$, $\tau(\mathbb{J}_2)$, $\tau(\mathbb{J}_3)$, $\tau(\mathbb{J}_4)$ and $\tau(\mathbb{J}_5)$) is listed as follows:

$$\theta! \text{ combinations for region labels : } \tau(\mathbb{J}_1)$$

$$\prod_{\kappa=0}^{15}\left(\frac{2^{\kappa}}{2}\right)! \text{ combinations for DC errors categories : } \tau(\mathbb{J}_2)$$

$$2^{\chi} \text{ combinations for AC sign information : } \tau(\mathbb{J}_3)$$

$$(M \times N)! \text{ combinations for block shuffling : } \tau(\mathbb{J}_4),$$

$$\prod_{i=1}^{M}\prod_{j=1}^{N}(n(i,j))! \text{ combinations for ZRV pairs : } \tau(\mathbb{J}_5)$$

where κ denotes the number of categories and $\chi = \sum_{i=1}^{M}\sum_{j=1}^{N} n(i,j)$. An adversary need to consider $\prod_{l=1}^{5}\tau(\mathbb{J}_l)$ difference combinations in total when performing the brute force attack. The number is significantly large. Specifically, when each block has at least one nonzero AC coefficient, here are at least 1024! combinations for $\tau(\mathbb{J}_4)$ itself for $M = N = 32$. Therefore, the proposed encryption method is robust against the brute force (i.e., plaintext only) attack.

Next, the chosen-plaintext attack scenario where the adversary has access to the corresponding encrypted image for any input image, is considered. The chosen-plaintext attack aims to determine the permutation and mapping operations by tracing the elements (e.g., coefficients) in the images before and after encryption. If it is infeasible to trace the elements of the input image and encrypted images by format-compliant selective encryption method, the encryption method is regarded as a robust against the chosen-plaintext attack. Note that, in the proposed format-compliant selective encryption method, the processes that handle DC and AC coefficients depend on the texture of the image. Specifically, the locations of DC coefficients are rearranged into regions deduced by the AC coefficients. In addition, the AC components in each block are shuffled in a manner such that (a) the nonzero AC coefficients are shuffled within each block, and (b) the AC blocks

are shuffled globally. Furthermore, the Huffman codewords for ZRV pairs can be bijectively mapped to other codewords with the same length, which reconstructs a different visual image. Hence, the aforementioned operations with the proposed format-compliant selective encryption method, which depends on the texture of the plaintext image, as well as the mapping of AC coefficients, are able to generate the huge combination number of DCT coefficients (DC and AC) between before and after encryption process. In other words, there is no straightforward way to trace the DCT coefficients in the corresponding plaintext and ciphertext images. Therefore, it is concluded that the proposed format-compliant selective encryption withstands the chosen-plaintext attack.

5.7.4 Discussions

The experiments were conducted on JPEG compressed grayscale images from the standard test images from USC-SIPI (USC-SIPI, 2014) and SIMPLIcity image dataset (Z. J. Wang et al., 2001). Color images from databases such as BSDS500 (Arbelàez et al., 2011) may be considered and they can be applied to the proposed encryption method by separately considering both luminance and chrominance channels (i.e., Y, Cb, and Cr). However, there is no commonly accepted quality metrics for color images, which have a similar reputation of either SSIM or PSNR for grayscale images. Therefore, the measured distortion caused by a format-compliant selective encryption may not be conclusive.

On the other hand, as an extension of application, the proposed encryption method can be applicable to the images compressed in other recent image compression form such as JPEG Extend Range (JPEG-XR), which includes three subbands coefficient values. However, the property of the three subbands may require analysis and redesign to meet the proposed four requirements in Section 5.3. In addition, the proposed encryption method also can be directly applicable to the I-frame in older generations of video compression standard (e.g., MPEG-1/2). However, manipulating motion compensated frames and

handling INTRA predicted blocks in recent standards such as H.264 and H.265 (H.264/5) require further analysis and effort to minimize bitstream size overhead while ensuring robustness against possible sketch attacks as well as classical cryptanalysis.

The aforementioned limitations and extensions will be pursued as future work.

5.8 Summary

In this Chapter, an encryption framework is proposed to design format-compliant selective encryption methods while aiming to minimize bitstream size overhead, which is a problem of current encryption methods. Specifically, a newly proposed sketch attack is considered to classify DC coefficients into some groups. The mechanism of these sketch attacks is utilized in the proposed framework to rearrange the DC coefficients for minimizing bitstream size overhead, which is also a problem of current encryption methods. Then, the proposed framework was augmented with various operations to form a complete format-compliant selective encryption method for JPEG images.

Experiment results verified that the proposed encryption method can severely degrade the quality of a JPEG compressed image while achieving smaller bitstream size overhead with more stable performance when compared to the conventional JPEG encryption methods. In addition, the proposed encryption method survives all sketch attacks considered, brute-force attack, as well as chosen-plaintext attack.

CHAPTER 6: TEXT DETECTION IN H.264/AVC COMPRESSED VIDEO

6.1 Overview

Multi-oriented text detection in compressed videos is a challenging task because the pixels are significantly de-correlated for efficient compression purpose. In this chapter, a novel fusion based multi-oriented text detection method in I-frame of H.264/AVC compressed video is proposed to demonstrate an application of the proposed sketch attacks in Chapter 2. The proposed text detection method utilizes four feature entities, each based on the sum of absolute value of judiciously selected AC coefficients. These extracted feature entities are fused into an image, which is, in turn, processed by Gaussian smoothing filter and clustered by *k*-means. Morphological operations are then applied to refine the output. Experiment results show that the proposed text detection method outperforms the conventional methods considered.

6.2 Introduction

As surveyed in Section 2.5, texts contained in a video include much information related the video, hence text detection is the most crucial step for video applications, i.e., video indexing and retrieval (J. Zhang & Kasturi, 2008; Jung et al., 2004). As of the writing of this thesis, there is only a paper (X. Qian et al., 2012) that proposed text detection method in H.264/AVC compressed video, which is the most successfully deployed video coding standard. However, this text detection method depends on threshold values for text block verification and it targets on graphics text of horizontal direction only but not scene text of arbitrary orientation.

Hence, in this chapter, a novel fusion method is proposed for detecting both graphics and scene multi-oriented text in H.264/AVC compressed video regardless of its scripts, font type and font size, which is a shift of paradigm from the conventional methods. The

proposed method utilizes four feature entities, each considering the sum of the absolute value of judiciously selected AC coefficients. These four feature entities are fused into an image to enhance edge contrast of the frame. The image is then classified into text and non-text regions by using *k*-means clustering. The text candidate regions are processed by morphological operations to remove noise. The processed image is considered to decide on the text regions. Experiments are carried out to verify the basic performance of the proposed method.

6.3 Proposed Fusion Based Multi-Oriented Text Detection

Qian et al. (X. Qian et al., 2012) only consider the sum of the absolute of AC coefficients and yet promising performance in text detection is achieved. In the same year, Roy et al. (Roy et al., 2012) enhance image contrast by proposing the multi-spectrum fusion method, which integrates wavelet and gradient information. Motivated by these methods, the fusion concept in (Roy et al., 2012) is improved for text detection in H.264/AVC compressed video by defining the feature entities directly in the compressed bitstream. The proposed fusion based text detection method consists of five parts, including (a) computing four feature entities; (b) fusing four feature entities into an image; (c) Gaussian filtering, (d) *k*-means clustering, and; (e) morphological operations. Figure 6.1 shows the process flow of the proposed method, where the input is transformed coefficients of each block in an H.264/AVC compressed video and the output is the detected text regions.

6.3.1 Motivation for Text Detection in H.264/AVC Video

Recall that sketch in Section 2.5.2 suggests that edge information can be extracted directly from the compressed domain. Results of the sketch images in Chapter 3 also suggest that sketch image is viable to represent the outline of the I-frame in H.264/AVC compressed video. In addition, video contents are generally stored the in compressed form, where the H.264/AVC standard is currently the most successfully deployed video coding standard.

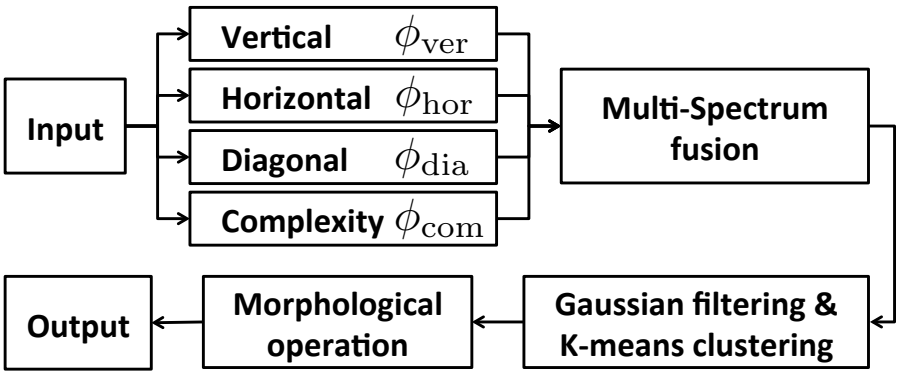


Figure 6.1: Flow diagram of the proposed text detection method

Since text in a video provide significant information about the video in question, it is justifiable to realize text detection directly in the compressed form without full decoding.

Furthermore, H.264/AVC performs IntDCT on the raw pixel and residual information, where IntDCT is also deployed in the latest video coding standard called HEVC (Sullivan et al., 2012). Hence, text detection in the IntDCT domain is a practical direction with potential future uses and is expected to reduce the computational cost because only partial decoding is required.

6.3.2 Four Feature Entities

Recall that $\mathbb{F}_{u,v}(i,j)$ is defined in Equation 2.1 as the (u,v)-th transformed and quantized coefficient value of the (i,j)-th $\gamma \times \gamma$ block in $\mathbb{T}$. In this chapter, the (u,v)-th absolute values of $\mathbb{F}_{u,v}(i,j)$ are considered to extract the four feature entities. The absolute value $\mathbb{A}_{u,v}(i,j)$ can be expressed as follows:

$$\mathbb{A}_{u,v}(i,j) = |\mathbb{F}_{u,v}(i,j)|. \tag{6.1}$$

In addition, the block size of 4×4 pixels is considered in the proposed method because this size is available in the BP (Wiegand et al., 2003). However, the higher frequency AC coefficients are usually of small magnitude due to quantization process. To utilize feature entries equally, normalization and rounding each quantity are considered.

First, the four feature quantities are defined as following sub-sections.

6.3.2.1 AC Coefficients with Vertical Information

The vertical entry (ϕ_{ver}) is defined by considering AC coefficients with vertical information as follows:

$$\phi_{\text{ver}}(i,j) = \text{round}\left(255 \times \frac{\lambda_V(i,j)}{\max\{\lambda_V(i,j)\}}\right), \tag{6.2}$$

where

$$\lambda_V(i,j) = \sum_{v=2}^{4} \mathbb{A}_{1,v}(i,j), \tag{6.3}$$

where round $(\cdot)$ rounds the real value to an 8-bit unsigned integer in the range of $[0,255]$ and $\lambda_V(i,j)$ is the sum of absolute vertical AC coefficients.

6.3.2.2 AC Coefficients with Horizontal Information

The horizontal entry (ϕ_{hor}) is defined by considering AC coefficients with horizontal information as follows:

$$\phi_{\text{hor}}(i,j) = \text{round}\left(255 \times \frac{\lambda_H(i,j)}{\max\{\lambda_H(i,j)\}}\right), \tag{6.4}$$

where the sum of absolute horizontal AC coefficients $\lambda_H(i,j)$ is defined as follows,

$$\lambda_H(i,j) = \sum_{u=2}^{4} \mathbb{A}_{u,1}(i,j). \tag{6.5}$$

6.3.2.3 AC Coefficients with Diagonal Information

The diagonal entry (ϕ_{dia}) is defined by considering AC coefficients with diagonal information as follows:

$$\phi_{\text{Dia}}(i,j) = \text{round}\left(255 \times \frac{\lambda_D(i,j)}{\max\{\lambda_D(i,j)\}}\right), \tag{6.6}$$

where the sum of absolute diagonal AC coefficients $\lambda_D(i,j)$ is defined as follows,

$$\lambda_D(i,j) = \sum_{u=2}^{3}\sum_{v=2}^{3} \mathbb{A}_{u,v}(i,j). \tag{6.7}$$

6.3.2.4 AC Coefficients with Complexity Information

The complex entry (ϕ_{com}) is defined by considering complexity measure as follows:

$$\phi_{\text{com}}(i,j) = \text{round}\left(255 \times \frac{\lambda_C(i,j)}{\max\{\lambda_C(i,j)\}}\right), \tag{6.8}$$

where the sum of absolute complexity AC coefficients $\lambda_C(i,j)$ is defined as follows,

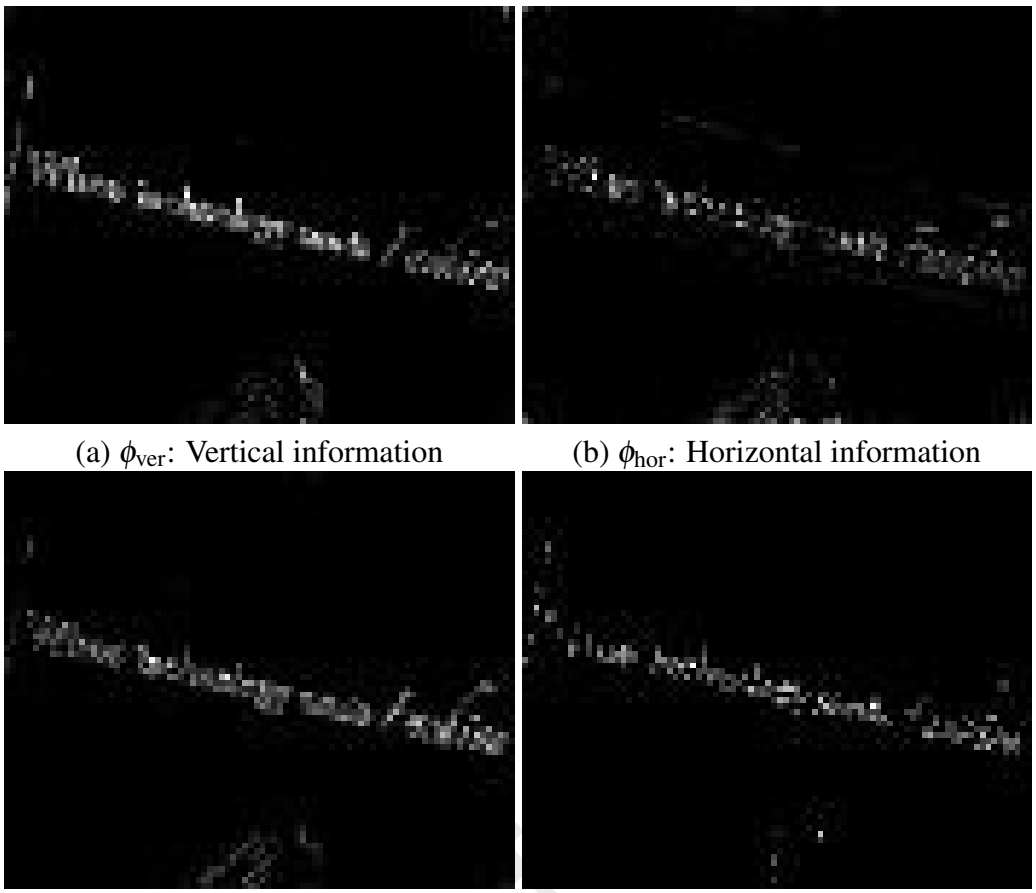
$$\lambda_C(i,j) = \sum_{u=2}^{3} \mathbb{A}_{4,v}(i,j) + \left(\sum_{v=2}^{3} \mathbb{A}_{u,4}(i,j)\right) + \mathbb{A}_{4,4}. \tag{6.9}$$

Figure 6.2 shows the images generated by the feature entities for an oriented text image.

6.3.3 Text Candidates Selection

To obtain a high contrast image for facilitating text detection, the maximum values among the four feature entities (viz., ϕ_{ver}, ϕ_{hor}, ϕ_{dia} and ϕ_{com}) are selected and fused into an image ϕ_T as follows:

$$\phi_T(i,j) = \max\{\phi_{ver}(i,j), \phi_{hor}(i,j), \phi_{dia}(i,j), \phi_{com}(i,j)\}. \tag{6.10}$$



(a) ϕ_{ver}: Vertical information (b) ϕ_{hor}: Horizontal information

(c) ϕ_{dia}: Diagonal information (d) ϕ_{com}: Complexity information

Figure 6.2: Four feature entities for oriented text

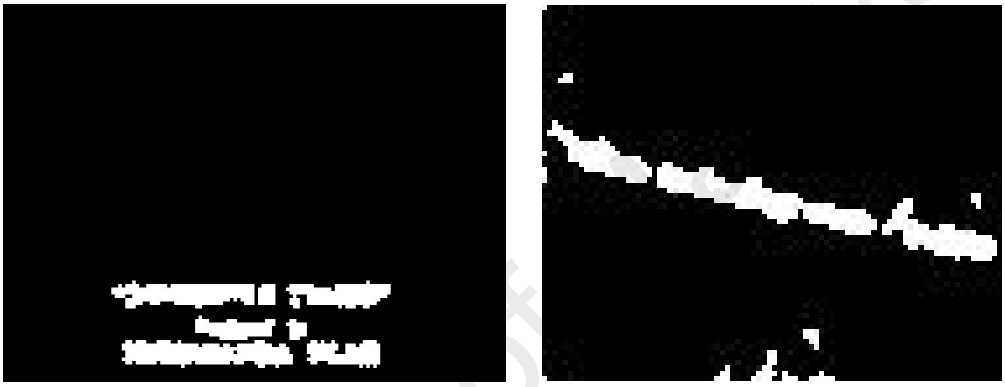
Figure 6.3(a) and (b) show the fused images for horizontal and oriented texts. Here, the fused images are first filtered by using Gaussian smoothing filter. Second, the text candidates are classified into the text and non-text regions. Then, morphological operations, specifically, 3×3 close and open filtering, are applied to the text regions. The refined text regions Ω_R are shown in Figure 6.4(a) and (b), where the clustered text regions are apparent for both horizontal and oriented texts.

6.4 Experiment Results and Discussions

The proposed text detection method is implemented in H.264/AVC using the reference software (Tourapis Michael et al., 2009) for partial decoding and Matlab for detecting text. Here, two datasets are considered, each consisting of graphics and/or scene text. Specifically, Dataset1 contains 100 images of sizes ranging from 352×240 to 816×448

(a) Horizontal (b) Oriented

Figure 6.3: Fused output for horizontal and oriented text



(a) Graphics horizontal (b) Scene oriented

Figure 6.4: Refined results

pixels, each with horizontal text. Dataset2 contains 100 images each of size 352×240 pixels with oriented text. These images are first encoded as an I-frame using the BP and decoded for detection purpose. In the following sub-sections, the results generated by the proposed fusion based method are presented.

6.4.1 Accuracy Performance

Two spatial domain methods (Shivakumara et al., 2010; C. Liu et al., 2005) and the latest H.264/AVC compressed domain method (X. Qian et al., 2012) are considered for comparison. Figure 6.5(a-d) show the output generated by the conventional methods and the proposed method for detecting horizontal text. It demonstrates that text lines are detected by all methods considered. Figure 6.6(a-d) shows an example of the output for oriented

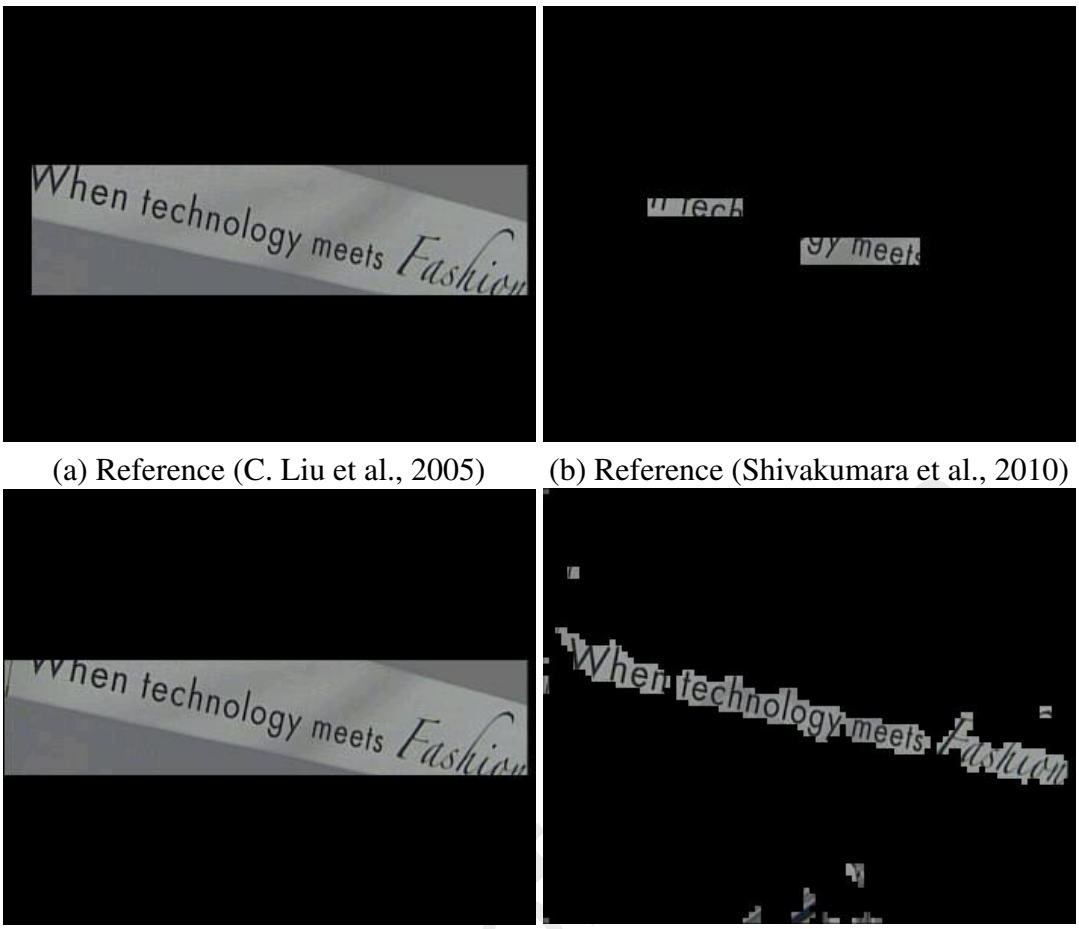
(a) Reference (C. Liu et al., 2005) (b) Reference (Shivakumara et al., 2010)

(c) Reference (X. Qian et al., 2012) (d) Proposed

Figure 6.5: Representative output for horizontal text

text. It is apparent that the existing methods failed to detect oriented text line because they are designed to detect horizontal text only, in which case the oriented text regions are removed as false positives or merged as horizontal text regions. Hence, some detected regions include false positives. On the other hand, the proposed method could detect the oriented text. To quantify the performance, manual counting of lines, i.e., correct and not correct, is performed, and then the resulting counted number is compared with the ground truth data. Here, three scores are considered, viz., *Precision* ("Number of correct lines" / "number of collected lines"), *Recall* ("Number of correct lines" / "ground truth"), and $F-measure$ ($2 \times Precision \times Recall/(Precision+Recall)$).

To further justify the proposed method, one more method utilizing only ϕ_{com} (complexity information) is considered. Table 6.1 compares the scores of each method for both image datasets. For dataset1, the scores of ϕ_{com} are less than those of the proposed



(a) Reference (C. Liu et al., 2005) (b) Reference (Shivakumara et al., 2010)

(c) Reference (X. Qian et al., 2012) (d) Proposed

Figure 6.6: Representative output for oriented text

method. However, for dataset2, the scores of ϕ_{com} dropped significantly because oriented text is assumed different features, which are not considered by ϕ_{com}. Results verify that the proposed method can exploit the feature entities in H.264/AVC for text detection. For Dataset1, References (Shivakumara et al., 2010; X. Qian et al., 2012) achieve higher precision than the proposed method, but suffer from low recall. These results suggest that the proposed method is more robust than the conventional methods in terms of text detection in video. For Dataset2, the existing methods suffer from very low recall, because the text candidates are removed by their refinement processes, i.e., bounding box size in References (Shivakumara et al., 2010; C. Liu et al., 2005) and projection in Reference (X. Qian et al., 2012). These results suggest that the proposed text detection method is superior in detecting oriented text while maintaining its performance for horizontal text.

Table 6.1: Scores for Precision, Recall, and False-Positive

		Spatial		Compressed		
Dataset	Score	C. Liu et al. (2005)	Shivakumara et al. (2010)	X. Qian et al. (2012)	ϕ_{com}	Proposed
dataset1	*Precision*	0.5292	0.8895	0.8244	0.7316	0.7681
	Recall	0.6128	0.6353	0.4060	0.5226	0.7594
	$F-measure$	0.5679	0.7412	0.5441	0.6096	0.7637
dataset2	*Precision*	0.2841	0.5483	0.3714	0.3993	0.6425
	Recall	0.2513	0.3410	0.1000	0.2949	0.7051
	$F-measure$	0.2667	0.4290	0.1576	0.3392	0.6724

6.4.2 Computation Complexity

In this section, computational complexity is discussed. For methods in the spatial domain, the total execution time from decoding the H.264/AVC compressed video to obtaining the detected text regions is collected. Table 6.2 compares the execution time of each method for both datasets, where *m* and *s* refer to minutes and seconds, respectively. Here, each method is executed on MacBook Pro (Retina, 13-inch, Mid 2014). As expected, the methods based on compressed domain yield less execution times than methods operating in the spatial domain. However, due to its complexity, i.e., computation of features and fusion parts, the proposed method executes slower than Qian's method (X. Qian et al., 2012). Nonetheless, it still executes faster than the considered spatial domain based methods.

Table 6.2: Computation time

	Spatial		Compressed	
Dataset	C. Liu et al. (2005)	Shivakumara et al. (2010)	X. Qian et al. (2012)	Proposed
Horizontal	33m29.2s	5m48.1s	1m42.5s	2m59.6s
Oriented	27m47.3s	4m55.5s	1m50.4s	1m46.4s

All in all, the proposed method is robust because it yields stable performance for detecting horizontal and oriented texts. Also, as mentioned in Section 6.4.1, the proposed method achieves higher accuracy than the latest method in the H.264/AVC compressed domain for the datasets considered. Furthermore, results suggest that the proposed method requires shorter execution time when compared to those in the spatial domain. Hence, the proposed text detection method is practical for text detection in H.264/AVC compressed video.

6.5 Summary

In this chapter, a novel fusion based text detection method is proposed to detect graphics and scene texts in H.264/AVC compressed videos. Text candidate lines are formed by fusing the proposed four feature entities computed directly from the AC coefficients. Results suggest that the proposed method outperforms the latest text detection method in the H.264/AVC compressed video for detecting graphics and scene texts of multi-orientations.

As future work, to improve precision, small false positive regions should be removed by considering entities in H.264/AVC compressed video, while suppressing the computational cost. Furthermore, exploring the performance, i.e., precision, recall, F-measure, on more datasets including text-frame and non-text frame, is necessary for practical applications. Implementation of real-time applications for text detection in H.264/AVC and HEVC compressed videos will be further pursued.

CHAPTER 7: CONCLUSION

In the last chapter, the research outcome of this study is summarized in the form of contributions made, particularly in feature extraction in the compressed domain. Next, since this study conducted under assumptions, therefore the limitations of this study are presented. Finally, future works and possible directions for applications by using the proposed extracted features in the compressed domain are discussed.

7.1 Summary

Five feature extraction methods, i.e., sketch attacks are proposed to generate outline information directly from block transform compressed images/videos. Then, an outline clearness assessment metric is put forward to evaluate the sketch images by considering both the global and local features. Next, a format-compliant selective encryption method for block transform images is presented, where the transformed coefficients are manipulated and processed based on the sketched image. Next, as an application of sketch image, text detection directly in the compressed video is demonstrated. Finally, the contributions, limitations, and future works, as well as possible directions of this study are presented.

7.2 Contributions

This study has achieved its objectives as follows:

1. Five sketch attacks, namely, DCEC, INCC, PLZ, SAC, and MBS, are put forward to extract an outline of video frame/image directly from H.264/AVC compressed videos and JPEG image encrypted by the current state-of-art format-compliant video/image encryption methods in Chapter 3.

2. A no-reference assessment metric is proposed by considering image entropy and spatial correlation to evaluate clearness of the outline image. Then, the obtained

sketch images are evaluated using the proposed assessment metric in Chapter 4

3. Recommendations on the design of format-compliant selective encryption for block transform based visual content are put forward. A rearrangement operation of DC coefficients, i.e., RDC, is proposed in Chapter 5. Next, a complete format-compliant selective encryption method for JPEG compressed image is formed by augmenting the proposed RDC with some conventional encryption operations to mask the AC coefficients.

4. The application of multi-oriented text detection is demonstrated by utilizing features from four proposed sketch attacks and the proposed fusion process in Chapter 6

7.3 Limitations

Although promising results were achieved, the proposed methods of this study have following limitations:

1. Generally, the resolution of a sketch image is much smaller than that of the original counterpart, and the sizes depend on the block transform window size and the size of macroblock size in a compression standard. The reduction in size of each sketch attack is reported in the second column of Table 7.1. In addition, since each sketch attack focus on a specific feature in the compressed domain, when the specific feature is significantly modified during encryption, a sketch attack may fail to sketch outline of the original video frame/image directly from the encrypted content. The third column of Table 7.1 records the corresponding features.

2. Literature survey carried out in this study suggests that there is no common agreement on ideal outline images and its evaluation, particularly for non-reference assessment. To address this problem, the proposed no-reference OCA metric is pro-

Table 7.1: Limitation of sketch attacks in compressed domain

Sketch image	Resolution	Exploited feature
ϕ_D	$1/\gamma$	The category of residue DC coefficients
ϕ_N	$1/\gamma$	The number of nonzero AC coefficients
ϕ_P	$1/\gamma$	The position of last nonzero AC coefficients
ϕ_S	$1/\gamma$	Sum of absolute AC coefficients
ϕ_E	$1/\gamma$	Modified sum of absolute AC coefficients
ϕ_T	$1/\gamma$	Fused four entities of selected AC coefficients
ϕ_B	$1/\delta$	The bits allocated to each MB

where γ is block transform window size and δ is the MB size.

posed, but it is based on empirical experiments and the observations. Therefore, the proposed no-reference OCA metric may be limited to the considered outline detectors and sketch attacks in this study.

3. The assumed encryption issues in this study, namely, bitstream size suppression and robust security against cryptanalysis and sketch attacks are solved by the proposed format-compliant encryption method. However, the proposed encryption framework may not be applicable to the newer/future compression standards, such as JPEG-XR, unless major redesigning takes place. It is because the newer compression standards employ adaptive entropy coding, where encryption may cause significant bitstream size overhead.

4. Although the proposed fusion based text detection method in Chapter 6 can detect multi-oriented text regions, it can only detect at the block size (i.e., restricted to the block size of 4×4), which is determined by the standard. Hence, it is a challenging task to outperform the conventional text detection methods in the non-compressed domain.

7.4 Future Works

Recall that, by utilizing some information about the used encryption method, appropriate sketch attack(s) can be launched to generate better sketch image(s) than just applying

sketch attack(s) one by one. More specifically, statistical analysis is conducted to obtain the information of better distribution of components, and then the component(s) with better distribution is(are) utilized for sketch attack.

In addition, the contributions of this study can be extended to design applications of multimedia content management tools, including, encryption, and text detection. Specifically, sketch image, which is a set of contour lines, can be utilized to realize application based on outline, e.g., human action recognition (Tom et al., 2015), moving object segmentation, etc.

Furthermore, assessment for the extracted features via sketch attack is still immature because this study proposes the first non-reference evaluation. A potential approach to justify the consistency of the proposed assessment metric is to perform automated image retrieval based on the extracted features using some well-labelled datasets.

Last but not least, the concept of sketch attack can be extended to the state-of-art block transform based compression technologies, including, JPEG-XR for still images, and HEVC for videos. Since the latest compression standards remove redundancy significantly, it is necessary to redesign sketch attack to efficiently and effectively extract the remaining information in the compressed contents (in the form of bitstream) as a feature extraction process in the compressed domain.

APPENDIX A: LIST OF PUBLICATIONS AND PAPERS PRESENTED

The following is the list of submitted / accepted journals and peer-viewed conference papers related to studies.

Journals:

[1] Kazuki Minemura, KokSheik Wong, C.-W Phan, Kiyoshi Tanaka. (2016). No-reference Clearness Assessment for Outline Images. Signal, Image and Video Processing. (submitted).

[2] Kazuki Minemura, KokSheik Wong, C.-W Phan, Kiyoshi Tanaka. (2016). A novel sketch attack for H.264/AVC format-compliant encrypted video. IEEE Transactions on Circuits and Systems for Video Technology. (Accepted).

[3] Kazuki Minemura, Koksheik Wong, Xiaojun Qi, and Kiyoshi Tanaka. (2016). A scrambling framework for block transform compressed image. Multimedia Tools and Applications (Accepted).

International Peer-Reviewed Conferences:

[1] Yiqi Tew, Kazuki Minemura, and Koksheik Wong. (2015). HEVC selective encryption using transform skip signal and sign bin. In IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. (pp. 963–970).

[2] Masaya Moriyama, Kazuki Minemura, and Koksheik Wong. (2015). Moving object detection in HEVC video by frame sub-sampling. In IEEE International Symposium on Intelligent Signal Processing and Communication Systems, (pp. 48–52).

[3] Kazuki Minemura and Koksheik Wong. (2014). Sketch attacks: A note on designing video encryption method in H.264/AVC. In IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. (pp. 1–7).

[4] Kazuki Minemura, Shivakumara Palaiahnakote, and Koksheik Wong. (2014). Multi-oriented text detection for intra-frame in H.264/AVC video. In IEEE Inter- national Sym-

posium on Intelligent Signal Processing and Communication Systems. (pp. 330–335).

[5] Kazuki Minemura, Shivakumara Palaiahnakote, and Koksheik Wong. (2014). A novel fusion based multi-oriented text detection method in intra-frame of H.264/AVC compressed video. In IIEEJ Image Electronics and Visual Computing Workshop.

[6] Sim Ying Ong, Kazuki Minemura, and Kok Sheik Wong. (2013). Progressive quality degradation in JPEG compressed image using DC block orientation with rewritable data embedding functionality. In IEEE International Conference on Image Processing, (pp. 4574–4578).

[7] Kazuki Minemura and KokSheik Wong. (2013). A rewritable data embedding in JPEG-XR compressed image with file size preservation. In H. B. Zaman, P. Robinson, P. Olivier, T. K. Shih, & S. Velastin (Eds.), Advances in Visual Informatics (Vol. 8237, pp. 569–580). Springer International Publishing.

[8] Kazuki Minemura and KokSheik Wong. (2012). Reversible visible watermark technique in DCT compressed domain. In IIEEJ Image Electronics and Visual Computing Workshop.

[9] Kazuki Minemura, Zahra Moayed, Koksheik Wong, Xiaojun Qi, and Kiyoshi Tanaka. (2012). JPEG image scrambling without expansion in bitstream size. In IEEE International Conference on Image Processing. (pp. 261–264).

REFERENCES

Ahn, J., Shim, H. J., Jeon, B., & Choi, I. (2004). Digital video scrambling method using intra prediction mode. In *Advances in Multimedia Information Processing - PCM 2004* (pp. 386–393). Springer Berlin Heidelberg.

Akhaee, M. A., Sahraeian, M. E., & Marvasti, F. (2010). Contourlet-based image watermarking using optimum detector in a noisy environment. *IEEE Transactions on Image Processing*, *19*(4), 967–80.

Arbelàez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 898–916.

Attar, A., Shahbahrami, A., & Rad, R. M. (2015). Image quality assessment using edge based features. *Multimedia Tools and Applications*, *75*(12), 7407–7422.

Bergeron, C., & Lamy-Bergot, C. (2005). Complaint selective encryption for H.264/AVC video streams. In *IEEE Workshop on Multimedia Signal Processing* (pp. 1–4).

Boho, A., Van Wallendael, G., Dooms, A., De Cock, J., Braeckman, G., Schelkens, P., … Van de Walle, R. (2013). End-to-end security for video distribution: The combination of encryption, watermarking, and video adaptation. *IEEE Signal Processing Magazine*, *30*(2), 97–107.

B.Pennebaker, W., & L.Mitchell, J. (1992). *JPEG: still image data compression standard*. Van Nostrand Reinhold.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(6), 679–698.

Chi, J., & Eramian, M. (2015). Enhancement of textural differences based on morphological component analysis. *IEEE Transactions on Image Processing*, *24*(9), 2671–2684.

Chunhua, L., Xinxin, Z., & Yuzhuo, Z. (2008). NAL level encryption for scalable video coding. In *Advances in Multimedia Information Processing - PCM 2008* (Vol. 5353, pp. 496–505).

Dey, B., & Kundu, M. K. (2013). Robust background subtraction for network surveillance in H.264 streaming video. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(10), 1695–1703.

Dimitrova, N., Agnihotri, L., Dorai, C., & Bolle, R. (2000). MPEG-7 videotext description scheme for superimposed text in images and video. *Signal Processing: Image Communication*, *16*(1), 137–155.

Dollar, P., & Zitnick, C. L. (2013). Structured forests for fast edge detection. In *IEEE International Conference on Computer Vision* (pp. 1841–1848).

Dubois, L., Puech, W., & Blanc-Talon, J. (2014). Smart selective encryption of H.264/AVC videos using confidentiality metrics. *annals of telecommunications - annales des télécommunications*, *69*(11–12), 569–583.

Feuvre, J. L., Thiesse, J., Parmentier, M., Raulet, M., & Daguet, C. (2014). Ultra high definition HEVC DASH data set. In *ACM Multimedia Systems Conference* (pp. 7–12).

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42.

Gonzalez, R. C., & Woods, R. E. (2006). *Digital image processing (3rd edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Goto, H. (2008). Redefining the DCT-based feature for scene text detection. *International Journal of Document Analysis and Recognition*, *11*(1), 1–8.

Grangetto, M., Magli, E., & Olmo, G. (2007). Conditional access to H.264/AVC video by means of redundant slices. In *IEEE International Conference on Image Processing* (pp. 485–488).

Group's, T. I. J. (2012). *The independent JPEG group's JPEG software.* Addison-Wesley Longman Publishing Co., Inc. Retrieved from `http://www.ijg.org/files/jpegsr8d.zip`

Gu, K., Zhai, G., Lin, W., Yang, X., & Zhang, W. (2015). No-reference image sharpness assessment in autoregressive parameter space. *IEEE Transactions on Image Processing*, *24*(10), 3218–3231.

Gu, K., Zhai, G., Yang, X., & Zhang, W. (2015). Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, *17*(1), 50–63.

Guan, J., Zhang, W., Gu, J., & Ren, H. (2015). No-reference blur assessment based on edge modeling. *Journal of Visual Communication and Image Representation*, *29*, 1–7.

Hou, X., Yuille, A., & Koch, C. (2013). Boundary detection benchmarking: Beyond f-measures. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2123–2130).

H.R. Sheikh, L. C., Z.Wang, & Bovik, A. (n.d.). *LIVE image quality assessment database release 2.* Retrieved from `http://live.ece.utexas.edu/research/quality`

ISO/IEC. (1994). *ISO/IEC 10918-1:1994 information technology – digital compression and coding of continuous-tone still images: requirements and guidelines.*

Jain, A., & Yu, B. (1998). Automatic text location in images and video frames. In *International Conference on Pattern Recognition* (Vol. 2, pp. 1497–1499).

Jiang, H., Liu, G., Qian, X., Nan, N., Guo, D., Li, Z., & Sun, L. (2008). A fast and effective text tracking in compressed video. In *IEEE International Symposium on Multimedia* (pp. 136–141).

Jiangtao, W., Severa, M., Wenjun, Z., Luttrell, M., & Weiyin, J. (2002). A format-compliant configurable encryption framework for access control of video. *IEEE Transactions on Circuits and Systems for Video Technology*, *12*(6), 545–557.

Johnson, M., Ishwar, P., Prabhakaran, V., Schonberg, D., & Ramchandran, K. (2004). On compressing encrypted data. *IEEE Transactions on Signal Processing*, *52*(10), 2992–3006.

Jung, K., Kim, K. I., & Jain, A. K. (2004). Text information extraction in images and video: a survey. *Pattern Recognition*, 977–997.

Kamble, V., & Bhurchandi, K. (2015). No-reference image quality assessment algorithms: A survey. *Optik - International Journal for Light and Electron Optics*, *126*(11-12), 1090–1097.

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Gomez i Bigorda, L., Robles Mestre, S., ... de las Heras, L.-P. (2013). ICDAR 2013 robust reading competition. In *International Conference on Document Analysis and Recognition* (pp. 1484–1493).

K.R.Rao, & P.Yip. (1990). *Discrete cosine transform: algorithms, advantages, applications*. San Diego, CA, USA: Academic Press Professional.

Kwon, S. G., Choi, W. I., & Jeon, B. (2005). Digital video scrambling using motion vector and slice relocation. In *Image Analysis and Recognition* (Vol. 3656, pp. 207–214).

Lee, H., & Nam, J. (2006). Low complexity controllable scrambler/descrambler for H.264/AVC in compressed domain. In *ACM International Conference on Multimedia.*

Li, H., & Jian Ren. (2007). A syntax aware error-tolerant encryption for secure multimedia communications. In *Workshop on Signal Processing Applications for Public Security and Forensics* (pp. 1–4).

Li, L., Lin, W., Wang, X., Yang, G., Bahrami, K., & Kot, A. C. (2015). No-reference image blur assessment based on discrete orthogonal moments. *IEEE Transactions on Cybernetics*, 1–1.

Li, Q., Han, Y., & Dang, J. (2016). Sketch4image: a novel framework for sketch-based image retrieval based on product quantization with coding residuals. *Multimedia Tools and Applications*, *75*(5), 2419–2434.

Li, W., & Yuan, Y. (2007). A leak and its remedy in JPEG image encryption. *International Journal of Computer Mathematics*, *84*(9), 1367–1378.

Li, Y., Wang, S., Tian, Q., & Ding, X. (2015). A survey of recent advances in visual feature detection. *Neurocomputing*, *149*, 736–751.

Lian, S., Liu, Z., Ren, Z., & Wang, H. (2007). Commutative encryption and watermarking in video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, *17*(6), 774–778.

Liang, J., & Yuen, S. Y. (2013). An edge detection with automatic scale selection approach to improve coherent visual attention model. *Pattern Recognition Letters*, *34*(13), 1519–1524.

Lin, W., & Jay Kuo, C.-C. (2011). Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, *22*(4), 297–312.

Liu, C., Wang, C., & Dai, R. (2005). Text detection in images based on unsupervised classification of edge-based features. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition* (pp. 610–614).

Liu, L., Liu, B., Huang, H., & Bovik, A. C. (2014). No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, *29*(8), 856–863.

Lopez-Molina, C., Baets, B. D., Bustince, H., Sanz, J., & Barrenechea, E. (2013). Knowledge-based systems multiscale edge detection based on gaussian smoothing and edge tracking. *Knowledge-Based Systems*, *44*, 101–111.

Lopez-Molina, C., De Baets, B., & Bustince, H. (2013). Quantitative error measures for edge detection. *Pattern Recognition*, *46*(4), 1125–1139.

Magli, E., Grangetto, M., & Olmo, G. (2006). Conditional access to H.264/AVC video with drift control. In *IEEE International Conference on Multimedia and Expo* (pp. 1353–1356).

Massoudi, A., Lefebvre, F., De Vleeschouwer, C., Macq, B., & Quisquater, J.-J. (2008). Overview on selective encryption of image and video: Challenges and perspectives. *EURASIP Journal on Information Security*, 1–18.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3–30.

McIlhagga, W. (2011). The canny edge detector revisited. *International Journal of Computer Vision*, *91*(3), 251–261.

Niu, X., Zhou, C., Ding, J., & Yang, B. (2008). JPEG encryption with file size preservation. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 308–311).

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66.

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2015). Visual privacy protection methods: A survey. *Expert Systems with Applications*, *42*(9), 4177–4195.

Parvin, Z., Seyedarabi, H., & Shamsi, M. (2014). A new secure and sensitive image encryption scheme based on new substitution with chaotic function. *Multimedia Tools and Applications*, 1–18.

Pazarci, M., & Dipcin, V. (2002). A MPEG2-transparent scrambling technique. *IEEE Transactions on Consumer Electronics*, *48*(2), 345–355.

Peng, F., wen Zhu, X., & Long, M. (2013). An ROI privacy protection scheme for H.264 video based on FMO and chaos. *IEEE Transactions on Information Forensics and Security*, *8*(10), 1688–1699.

Podesser, M., Schmidt, H.-P., & Uhl, A. (2002). Selective bitplane encryption for secure transmission of image data in mobile environments. In *IEEE Nordic Signal Processing Symposium* (pp. 4–6).

Qian, F., Guo, J., Sun, T., & Wang, T. (2015). Quantitative assessment of laser-dazzling effects through wavelet-weighted multi-scale SSIM measurements. *Optics & Laser Technology*, *67*, 183–191.

Qian, J., Wu, D., Li, L., Cheng, D., & Wang, X. (2014). Image quality assessment based on multi-scale representation of structure. *Digital Signal Processing*, *33*, 125–133.

Qian, X., Liu, G., Wang, H., & Su, R. (2007). Text detection, localization, and tracking in compressed video. *Signal Processing: Image Communication*, *22*(9), 752–768.

Qian, X., Wang, H., & Hou, X. (2012). Video text detection and localization in intra-frames of H.264/AVC compressed video. *Multimedia Tools and Applications*, 1–16.

Robert, M. H., & Linda, G. S. (1992). *Computer and robot vision* (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Roy, S., Shivakumara, P., Roy, P. P., & Tan, C. L. (2012). Wavelet-Gradient-Fusion for Video Text Binarization. *International Conference on Pattern Recognition*(Icpr), 3300–3303.

Shahid, Z., Chaumont, M., & Puech, W. (2009). Fast protection of H.264/AVC by selective encryption of CABAC. In *IEEE International Conference on Multimedia and Expo* (pp. 1038–1041).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shen, H., Zhuo, L., & Zhao, Y. (2014). An efficient motion reference structure based selective encryption algorithm for H.264 videos. *IET Information Security*, *8*(3), 199–206.

Shiguo, L., Jinsheng, S., Guangjie, L., & Zhiquan, W. (2008). Efficient video encryption scheme based on advanced video coding. *Multimedia Tools and Applications*, *38*(1), 75–89.

Shiguo, L., Jinsheng, S., & Zhiquan, W. (2004). A novel image encryption scheme based-on JPEG encoding. In *IEEE Symposium on Information Visualization* (pp. 217–220).

Shiguo, L., Zhongxuan, L., Zhen, R., & Haila, W. (2006). Secure advanced video coding based on selective encryption algorithms. *IEEE Transactions on Consumer Electronics*, *52*(2), 621–629.

Shiguo, L., Zhongxuan, L., Zhen, R., & Zhiquan, W. (2005). Selective video encryption based on advanced video coding. In *Advances in Multimedia Information Processing - PCM 2005* (pp. 281–290).

Shivakumara, P., Phan, T. Q., & Tan, C. (2010). New wavelet and color features for text detection in video. In *International Conference on Pattern Recognition* (pp. 3996–3999).

Shivakumara, P., Phan, T. Q., & Tan, C. L. (2011). A laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 412–419.

Shivakumara, P., Sreedhar, R., Trung, Q. P., Shijian, L., & Tan, C. (2012). Multioriented video scene text detection through bayesian classification and boundary growing. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(8), 1227–1235.

Shivakumara, P., Trung, Q. P., Shijian, L., & Chew, L. T. (2013). Gradient vector flow and grouping-based method for arbitrarily oriented scene text detection in video images. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(10), 1729–1739.

Siu, Y., Kei Au, Shuyuan, Z., & Bing, Z. (2009). Partial video encryption based on alternating transforms. *IEEE Signal Processing Letters*, *16*(10), 893–896.

Socek, D., Kalva, H., Magliveras, S. S., Marques, O., Culibrk, D., & Furht, B. (2007). New approaches to encryption and steganography for digital videos. *Multimedia Systems*, *13*(3), 191–204.

Spinsante, S., Chiaraluce, F., & Gambi, E. (2005). Masking video information by partial encryption of H.264/AVC coding parameters. In *European Signal Processing Conference* (pp. 1–4).

Stutz, T., & Uhl, A. (2012). A survey of H.264 AVC/SVC encryption. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(3), 325–339.

Su, P., Hsu, C., & Wu, C. (2011). A practical design of content protection for H.264/AVC compressed videos by selective encryption and fingerprinting. *Multimedia Tools and Applications*, *52*(2-3), 529–549.

Subramanyam, A. V., & Emmanuel, S. (2014). Partially compressed-encrypted domain robust JPEG image watermarking. *Multimedia Tools and Applications*, *71*(3), 1311–1331.

Sullivan, G. J., Ohm, J.-r., Han, W.-j., & Wiegand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(12), 1649–1668.

Takayama, M., Tanaka, K., Takagi, K., & Nakajima, Y. (2008). A scalable video scrambling method in MPEG compressed domain. In *International Symposium on Communications, Control and Signal Processing* (pp. 1035–1040).

Takayama, M., Tanaka, K., Yoneyama, A., & Nakajima, Y. (2006). A video scrambling scheme applicable to local region without data expansion. In *IEEE International Conference on Multimedia and Expo* (pp. 1349–1352).

Tang, Z., Zhang, X., & Lan, W. (2015). Efficient image encryption with block shuffling and chaotic map. *Multimedia Tools and Applications*, *74*(15), 5429–5448.

Thomas, N., Lefol, D., Bull, D., & Redmill, D. (2007). A novel secure H.264 transcoder using selective encryption. In *IEEE International Conference on Image Processing* (Vol. 4, pp. 85–88).

Tom, M., Babu, R. V., & Praveen, R. G. (2015). Compressed domain human action recognition in H.264/AVC video streams. *Multimedia Tools and Applications*, *74*(21), 9323–9338.

Tong, C. S., Zhang, Y., & Zheng, N. (2005). Variational image binarization and its multiscale realizations. *Journal of Mathematical Imaging and Vision*, *23*(2), 185–198.

Tourapis Michael, A., Sühring, K., , & Sullivan, G. (2009). H.264/14496-10 AVC reference software manual [Computer software manual].

USC-SIPI. (2014). *Signal and image processing institute: USC-SIPI image database.* Retrieved from `http://sipi.usc.edu/database/misc.zip`

Van Droogenbroeck, M., & Benedett, R. (2002). Techniques for a selective encryption of uncompressed and compressed images. *Advanced Concepts for Intelligent Vision Systems*, 90–97.

Wang, Y., O'Neill, M., & Kurugollu, F. (2013). Privacy region protection for H.264/AVC by encrypting the intra prediction modes without drift error in i frames. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 2964–2968).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Z., Simoncelli, E., & Bovik, A. (2003). Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems, and Computers* (Vol. 2, pp. 1398–1402).

Wang, Z. J., Li, J., & Wiederhold, G. (2001). SIMPLIcity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(9), 947–963.

Wiegand, T., Sullivan, G., Bjontegaard, G., & Luthra, A. (2003). Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(7), 560–576.

Won, Y. G., Bae, T. M., & Ro, Y. M. (2006). Scalable protection and access control in full scalable video coding. In *Digital Watermarking* (pp. 407–421).

Wong, K., & Tanaka, K. (2010). DCT based scalable scrambling method with reversible data hiding functionality. In *International Symposium on Communications, Control, and Signal Processing* (pp. 1–4).

Wu, C.-P., & Kuo, C.-C. (2005). Design of integrated multimedia compression and encryption systems. *IEEE Transactions on Multimedia*, *7*(5), 828–839.

Wu, Q., Li, H., Meng, F., Ngan, K. N., & Zhu, S. (2015). No reference image quality assessment metric via multi-domain structural information and piecewise regression. *Journal of Visual Communication and Image Representation*, *32*, 205–216.

Xiph.org video test media. (n.d.). Retrieved from `http://media.xiph.org/video/derf/`

Yang, L., Chun, Y., & Yuzhuo, Z. (2007). A new digital rights management system in mobile applications using H.264 encryption. In *International Conference on Advanced Communication Technology* (Vol. 1, pp. 583–586).

Yeongyun, K., Sung, H. J., Tae, M. B., & Yong, M. R. (2007). A selective video encryption for the region of interest in scalable video coding. In *IEEE Region 10 Conference* (pp. 1–4).

Yi, S., Labate, D., Easley, G. R., & Krim, H. (2009). A shearlet approach to edge analysis and detection. *IEEE Transactions on Image Processing*, *18*(5), 929–941.

Yongsheng, W., O'Neill, M., & Kurugollu, F. (2013). A tunable encryption scheme and analysis of fast selective encryption for CAVLC and CABAC in H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(9), 1476–1490.

Yuan, L., Liwei, L., Zhaopin, S., & Jianguo, J. (2005). A new video encryption algorithm for H.264. In *International Conference on Information Communications & Signal Processing* (pp. 1121–1124).

Zeng, B., Yeung, S.-K. A., Zhu, S., & Gabbouj, M. (2014). Perceptual encryption of H.264 videos: Embedding sign-flips into the integer-based transforms. *IEEE Transactions on Information Forensics and Security*, *9*(2), 309–320.

Zeng, W., & Lei, S. (2003). Efficient frequency domain selective scrambling of digital video. *IEEE Transactions on Multimedia*, *5*(1), 118–129.

Zhang, J., & Kasturi, R. (2008). Extraction of text objects in video documents: Recent progress. In *International Workshop on Document Analysis Systems* (pp. 5–17).

Zhang, M., & Tong, X. (2014). A new chaotic map based image encryption schemes for several image formats. *Journal of Systems and Software*, *98*, 140–154.

Zhong, Y., Zhang, H.-J., & Jain, A. (1999). Automatic caption localization in compressed video. In *International Conference on Image Processing* (Vol. 2, pp. 96–100).

Zhou, J., Liu, X., Au, O. C., & Tang, Y. Y. (2014). Designing an efficient image encryption-then-compression system via prediction error clustering and random permutation. *IEEE Transactions on Information Forensics and Security*, *9*(1), 39–50.